Active Online Learning for Social Media Analysis to Support Crisis Management

Daniela Pohl, Abdelhamid Bouchachia *SMIEEE*, and Hermann Hellwagner *SMIEEE*

Abstract—People use social media (SM) to describe and discuss different situations they are involved in, like crises. It is therefore worthwhile to exploit SM contents to support crisis management, in particular by revealing useful and unknown information about the crises in real-time. Hence, we propose a novel active online multiple-prototype classifier, called AOMPC. It identifies relevant data related to a crisis. AOMPC is an online learning algorithm that operates on data streams and which is equipped with active learning mechanisms to actively query the label of ambiguous unlabeled data. The number of queries is controlled by a fixed budget strategy. Typically, AOMPC accommodates partly labeled data streams. AOMPC was evaluated using two types of data: (1) synthetic data and (2) SM data from Twitter related to two crises, Colorado Floods and Australia Bushfires. To provide a thorough evaluation, a whole set of known metrics was used to study the quality of the results. Moreover, a sensitivity analysis was conducted to show the effect of AOMPC's parameters on the accuracy of the results. A comparative study of AOMPC against other available online learning algorithms was performed. The experiments showed very good behavior of AOMPC for dealing with evolving, partly-labeled data streams.

Index Terms—Online Learning, Multiple Prototype Classification, Active Learning, Social Media, Crisis Management

1 Introduction

The primary task of crisis management is to identify specific actions that need to be carried out before (prevention, preparedness), during (response), and after (recovery and mitigation) a crisis occurred [28]. In order to execute these tasks efficiently, it is helpful to use data from various sources including the public as witnesses of emergency events. Such data would enable emergency operations centers to act and organize the rescue and response. In recent years, a number of research studies [47] have investigated the use of social media as a source of information for efficient crisis management. A selection of such studies, among others, encompasses Norway Attacks [45], Minneapolis Bridge Collapse [35], California Wildfire [62], Colorado Floods [18], and Australia Bushfires [23], [22]. The extensive use of SM by people forces (re)thinking the public engagement in crisis management regarding the new available technologies and resulting opportunities [13].

Our previous work on SM in emergency response focused on offline and online clustering of SM messages. The offline clustering approach [48] was applied to identify sub-events (specific hotspots) from SM data of a crisis for an after-the-fact analysis. Online clustering [46] was used to identify sub-events that evolve over time in a dynamic way. In particular, online feature selection mechanisms were devised as well, so that SM data streams can be accommodated continuously and incrementally.

It is interesting to note that people from emergency departments (e.g., police forces) already use SM to gather, monitor, and to disseminate information to inform the public [21]. Hence, we propose a learning algorithm, AOMPC, that relies on active learning to accommodate the user's feedback upon querying the item being processed. Since AOMPC is a classifier, the query is related to labeling that item.

The primary goal in using user-generated contents of SM is to discriminate valuable information from irrelevant one. We propose classification as the discrimination method. The classifier plays the role of a filtering machinery. With the help of the user, it recognizes the important SM items (e.g., tweets), that are related to the *event* of interest. The selected items are used as cues to identify *sub-events*. Note that an *event* is the crisis as such, while *sub-events* are the topics commonly discussed (i.e., hotspots like flooding, collapsing of bridges, etc. in a specific area of a city) during a crisis. These sub-events can be identified by aggregating the messages posted on SM networks describing the same specific topic [46], [49].

We propose a *Learning Vector Quantization* (LVQ)-like approach based on multiple prototype classification. The classifier operates *online* to deal with the *evolving stream of data*. The algorithm, named *active online multiple prototype classifier* (*AOMPC*), uses unlabeled and labeled data which are tagged through active learning. Data items which fall into ambiguous

- *Daniela Pohl and Hermann Hellwagner are with the Institute of Information Technology, Alpen-Adria-Universität Klagenfurt, 9020 Klagenfurt, Universitätsstr. 65-67, Austria. (phone: +43 463 2700 3688, fax: +43 463 2700 993688, e-mail: {daniela,hellwagn}@itec.aau.at)*
- *Abdelhamid Bouchachia is with the Smart Technology Research Centre, Bournemouth University, Poole, BH12 5BB, UK (phone: +44 (0) 1202 962401, fax: +44 (0) 1202 965314, e-mail: abouchachia@bournemouth.ac.uk)*

regions are selected for labeling by the user. The number of queries is controlled by a budget. The requested items help to direct the AOMPC classifier to a better discriminatory capability. While AOMPC can be applied to any streaming data, here we consider in particular SM data.

The contributions of this paper are as follows:

- An original online learning algorithm, AOMPC, is proposed to handle data streams in an efficient way. It is a multi-prototype LVQ-like algorithm inspired by our previous work [9], [8].
- As part of AOMPC, an active learning strategy is introduced to guide AOMPC towards accurate classification, and in this paper towards subevent detection. Such a strategy makes use of budget and uncertainty notions to decide when and what to label.
- AOMPC is evaluated on different data: synthetic datasets (synthetic numerical data, generated microblogs, which are geo-tagged) and real-world datasets collected from Twitter related to two crises, Colorado Floods in 2013 and Australia Bushfires in 2013. The choice and the use of all these datasets was motivated by their diversity. That allows to thoroughly evaluate AOMPC because these datasets have different characteristics.
- A sensitivity analysis based on the different AOMPC parameters and datasets is carried out.
- A comparison of AOMPC against well-known online algorithms is conducted and discussed.

The paper has the following structure. Section 2 presents the related work covering streaming and SM analysis. Section 3 introduces the classification algorithm and describes the processing steps, including the active learning facets. Section 4 discusses the empirical evaluation of AOMPC after describing the datasets used. Section 5 concludes the paper.

2 Related Work

The problem addressed in this paper is related to several topics: multiple prototype and Learning Vector Quantization (LVQ) classification, online learning for classification, active learning with budget planning, and social media analysis (i.e., natural language processing). A short overview of these topics is presented in the following.

2.1 Multiple Prototype Classification and LVQ Classification

A prototype-based classification approach operates on data items mapped to a vector representation (e.g., vector space model for text data). Data points are classified via prototypes considering similarity measures. Prototypes are adapted based on items related/similar to them.

A Rocchio classifier [37] is an example of a single prototype-based classifier. It distinguishes between two classes, e.g., “relevant” and “irrelevant”. In real world-scenarios, due to the nature of the data, it is often not possible to describe the data with a single prototype-based classifier. Multiple prototype classifiers (i.e., several prototypes) are needed.

Self organizing maps (SOM) introduced by Kohonen [32] are an unsupervised version of prototype-based classification, also known as LVQ. In this case, prototypes are initialized (e.g., randomized) and adapted. SOM was also used for SM analysis in the context of crisis management to identify important hotspots [48].

LVQ has been applied to several areas, e.g., robotics, pattern recognition, image processing, text classification etc. [20], [32], [60]. LVQ - in the context of similarity representation, rather then vector-based representation - is analyzed by Hammer et al. [25]. Mokbel et al. [38] describe an approach to learn metrics for different LVQ classification tasks. They suggest a metric adaptation strategy to automatically adapt metric parameters.

Bezdek et al. [6] review several offline multiple prototype classifiers, e.g., LVQ, fuzzy LVQ, and the deterministic Dog-Rabbit (DR) model. The latter limits the movement of prototypes and is similar to our approach. However, in contrast to our approach, DR uses offline adaptation of the learning rate. The time-based learning rate of our algorithm considers concept drift (i.e., changes of the incoming data) directly during the update of the prototypes.

In contrast to the previous approaches, Bouchachia [8] proposes an incremental supervised LVQ-like competitive algorithm that operates online. It consists of two stages. In the first stage (learning stage), the notions of winner reinforcement and rival repulsion are applied to update the weights of the prototypes. In the second stage (control stage), two mechanisms, *staleness* and *dispersion* are used to get rid of dead and redundant prototypes.

A summary of different prototype based learning approaches can be found in Biehl et al. [7]. In this study, we deal with online real-time classification and we propose a multi-prototype quantization algorithm, where the winning prototype is adapted based on the input. In particular, the algorithm relies on online learning and active learning.

2.2 Online Learning and Active Learning (with Budget Planning)

Online learning receives data items in a continuous sequence and processes them once to classify them accordingly [64]. Bouchachia and Vanaret [10], [11] use Growing Gaussian Mixture Models for online classification. Compared to the algorithm proposed in this work, there is a difference in adapting the learning rate and representing the prototypes. Reuter et al. [52] use multiple prototypes representing an event. New incoming items are assigned to the most similar events (by using an offline-trained SVM) or otherwise new events are created.

Another important topic in streaming analysis is active learning to improve results of classification with an amount of labeled data actively asked by the system [55]. Ienco et al. [29] use a pre-clustering step to identify relevant items to be labeled by the user. In Smailović et al. [57] active learning is used to improve the sentiment analysis of incoming tweets as an indicator for stock movements. Hao et al. [27] design two active learning algorithms (Active Exponentially Weighted Average Forecaster and Active Greedy Forecaster) which includes feedback of experts for labeling. The approach considers confidence of labels from the classifier compared to a set of experts. Hao et al. [26] also introduce online active learning considering second order information, e.g. based on covariance matrix. Ma et al. [36] combine decision trees with active learning. This approach improves the learning step for decision trees. Bouguelia et al. [12] use instance weighting for active online learning. They consider the weight that must be changed to cause the classifier changing its prediction. If only a small change in weight changes the original classification, then the classifier is highest uncertain about the item.

Monzafari et al. [39] study different batch-based active learning approaches and define two uncertainty strategies to query labels from crowdsourcing platforms. In addition, the authors also define a budget or goal constraint to limit labeling. Žliobaitė et al. [63] use active learning combined with streaming data. They suggest several processing mechanisms to identify uncertainty regions especially for handling data drifts. It is also important to minimize the number of queries, asking an expert for labels. Žliobaitė et al. [63] include a moving average over the incoming items and the amount of already labeled items to estimate the budget. We adopted this mechanism together with the uncertainty strategies.

Based on categorization of active learning approaches by Settles et al. [55], our implementation is classified as a stream-based selective sampling approach, considering different strategies to request instances for labeling. In addition, we use an online feature selection approach described later.

2.3 Social Media Analysis for Crisis Management

Recent research studies SM from several technical perspectives. Due to space limitations, we describe existing SM analysis frameworks mostly in the context of crisis management, although there are several frameworks in other contexts, e.g., Twitterbeat [56] and HarVis [2]. Backfried et al. [3] describe an analysis approach based on visual analytics for combining information from different sources with a specific focus on multilingual issues. Vieweg and Hodges [30], [61] describe the Artificial Intelligence for Disaster Response (AIDR) platform, where persons annotate incoming tweets (similar to Amazon Mechanical Turk). The tweets are then used to train classifiers to identify more relevant tweets. AIDR allows to classify incoming tweets based on different information categories, e.g., damage report, casualties, advises, etc. Chen et al. [15] analyse tweets related to Flu to identify topics for predicting the Flu-peak. Neppalli et al. [40] perform sentiment analysis based on social media related to Hurricane Sandy. The work shows that sentiment of users is related to the distance of the Hurricane to the users. Twitcident described by Abel et al. [1] is a framework to search and filter Twitter messages through specific profiles (e.g., keywords). Terpstra et al. [59] show the usage of Twitcident in crisis management. Tweak-the-Tweet introduced by Starbird et al. [58] defines a grammar which can be easily integrated in tweets and therefore automatically parsed. Also, TEDAS described by Li et al. [34] is a system to detect high-level events (e.g., all car accidents in a certain time period) using spatial and temporal information. Yin et al. [66], [65] design a situational awareness platform for SM. Tweets are analyzed based on bursty keywords to identify emergent incidents. Ragini et al. [50] combine several techniques to identify people in danger. They examined rule based classification and several machine learning approaches, like SVM, for hybrid classification.

Additional information on social media analysis in different crises can be found in Reuter and Kaufhold [51]. Due to the importance of SM, it is our aim to support emergency management when using the content of SM platforms. Currently, there are systems with crowd-sourcing platform characteristics, but no procedure (like active learning) is available to directly involve emergency management personnel in filtering relevant information.

3 Active Online Multiple Prototype Classifier (AOMPC)

Due to the fact that SM data is noisy, it is important to identify relevant SM items for the crisis situation at hand. The idea is to find an algorithm that performs this classification and also handles ambiguous items in a reasonable way. Ambiguous denotes items where a clear classification is not possible based on the current knowledge of the classifier. The knowledge should be gained by asking an expert for feedback. The algorithm should be highly self-dependent, by asking the expert only labels for a limited number of items.

Therefore, we propose an original approach that combines different aspects - such as online learning and active learning - to build a hybrid classifier, AOMPC. AOMPC learns from both, labeled and

TABLE 1
List of symbols used

Variable	Description
x	Input (one item) received by the data stream X with bt_{CT} batches
V	Set of currently known prototypes
α	A parameter used in Alg. 1 to compute the staleness of a prototype. It is given as: $\alpha = e^{\frac{-log2}{\beta}}$, where β is the half-life span, denoted hereafter as (1/2)-life-span, described in [31] that refers to the amount of time required for a quantity to fall to half its value as measured at the beginning of the time period.
I	Set of indices i indicating the prototypes $\mathbf{v}_i$
$dist$	Appropriate distance measure; see Algorithm 2
UT	Threshold used to identify uncertainty
CT	Current time
LTU	Last time the prototype was updated (i.e., the winner)
S	List of nearest prototypes in ascending order to the current input **x**
$label$	Labels are: $relevant$, $irrelevant$, and $unknown$

Algorithm 1 : Steps of AOMPC

```
Input: Data stream $X$
Output: List of prototypes $V$
 1: CT=1; LTU=CT;
 2: Let CT and LTU indicate the current time and the last time a prototype was updated respectively
 3: for batch $bt_{CT}$ of $X$ do
 4:   for incoming input $\mathbf{x}$ of $bt_{CT}$ do
 5:     Compute distance $\varphi_i$ between $\mathbf{x}$ and all prototypes $\mathbf{v}_i$, $i = 1 \cdots |V| = I$, as follows:
          if $(inaction(\mathbf{v}_i) > 0)$ $\varphi_i = inaction(\mathbf{v}_i) \cdot dist(\mathbf{v}_i, \mathbf{x})$
          else $\varphi_i = dist(\mathbf{v}_i, \mathbf{x})$ end if        (1)
        such that $inaction(\mathbf{v}_i) = 1 - \alpha^{(CT - \mathbf{v}_i.LTU)}$
 6:     Compute list of nearest prototypes $S$ based on sorted index $I$ such that $S = createSortedList(I,\ (x,y)) : (\varphi_x \leq \varphi_y)$
 7:     check = uncertainty(x) and within_budget();
 8:     if check = true then
 9:       Query the label of x
10:     else
11:       $\mathbf{x}.label = unknown$
12:     end if
13:     if $S \neq \{\}$ then
14:       Let $j$ be the index of the closest prototype: $j = S(1)$
15:       if $\mathbf{x}.label = unknown$ then
16:         Assign the data item to $\mathbf{v}_j$
17:       else
18:         if $\mathbf{x}.label = \mathbf{v}_j.label$ then
19:           Reinforce $\mathbf{v}_j$ with x using only the common features:
                $\mathbf{v}_j = \mathbf{v}_j + \alpha^{CT-LTU}(\mathbf{x} - \mathbf{v}_j)$
20:           Add the non-common features of x to $\mathbf{v}_j$: $\mathbf{v}_j.feature = \alpha^{CT-LTU}$(x.feature)
21:         else
22:           Go to line 26
23:         end if
24:       end if
25:     else
26:       Initialize a new prototype: $\mathbf{v}_{new}$=x
27:       $\mathbf{v}_{new}.label = \mathbf{x}.label$; $\mathbf{v}_{new}.LTU = CT$
28:       $V = V \cup \{\mathbf{v}_{new}\}$
29:     end if
30:   end for
31:   Update winning clusters in $bt_{CT}$ with $LTU = CT$
32:   CT = CT + 1;
33: end for
```

unlabeled data, in a continuous and evolving way. In this context, AOMPC is designed to distinguish between relevant and irrelevant SM data related to a crisis situation in order to identify the needs of individuals affected by the crisis. AOMPC relies on active learning. It implies the intervention of a user in some situations to enhance its effectiveness in terms of identifying relevant data and the related event in the SM stream of data (see Fig. 1). The user is asked to label an item if there is a high uncertainty about the classification as to whether it is relevant or irrelevant. The classifier assigns then the item (be it actively labeled or unlabeled) to the closest cluster or uses it to create a new cluster. A cluster - in this case - represents either relevant (i.e., specific information about the crisis of interest) or irrelevant information (i.e., not related to the crisis). The process flow and the steps of AOMPC are shown in Fig. 1. AOPMC is described in Algorithm 1. The used symbols are defined in Tab. 1. CT and LTU are updated in batch-mode due to the feature selection method used (see Section 3.3 for details). The algorithm could also be used in item-wise mode. The general idea of this algorithm is that the longer a prototype is stale (not updated), the slower it should move to a new position. The learning rate α is a function of the last time the prototype was a *winner* (i.e., α can be seen as a *forgetting factor*). The winning prototype is computed based on the learning rate (steps 5-6). If there is an uncertainty detected (see Section 3.2) and enough budget is available (see Section 3.1), the label is queried (steps 7-11). Otherwise (e.g., not enough budget) the winning prototype defines the label (step 16). When a prototype wins the competition among all other neighboring prototypes based on the queried label, it is updated to move in the direction of the new incoming item (steps 17-20). In case the new input comes with new features, the prototype's feature vector is extended to cover those new textual features (see step 20). In general, AOMPC is capable of accommodating new features. In the case of textual input, like in this study, the evolution of the vocabulary over time is captured. When no prototype is sufficiently close to the new item (step 22), a new prototype is created to accommodate that item (steps 26-28).

Algorithm 1 relies on the computation of the distance between the input and the existing prototypes (e.g., Euclidean distance in Algorithm 2). Because the SM items usually consist of a textual description (c.f., tweets), we apply the Jaccard coefficient [37] as a text-based distance ($dist\_text$) (see Algorithm 2, steps 2-3). If the social media items consist of two parts, the body of the message and the geo-location that

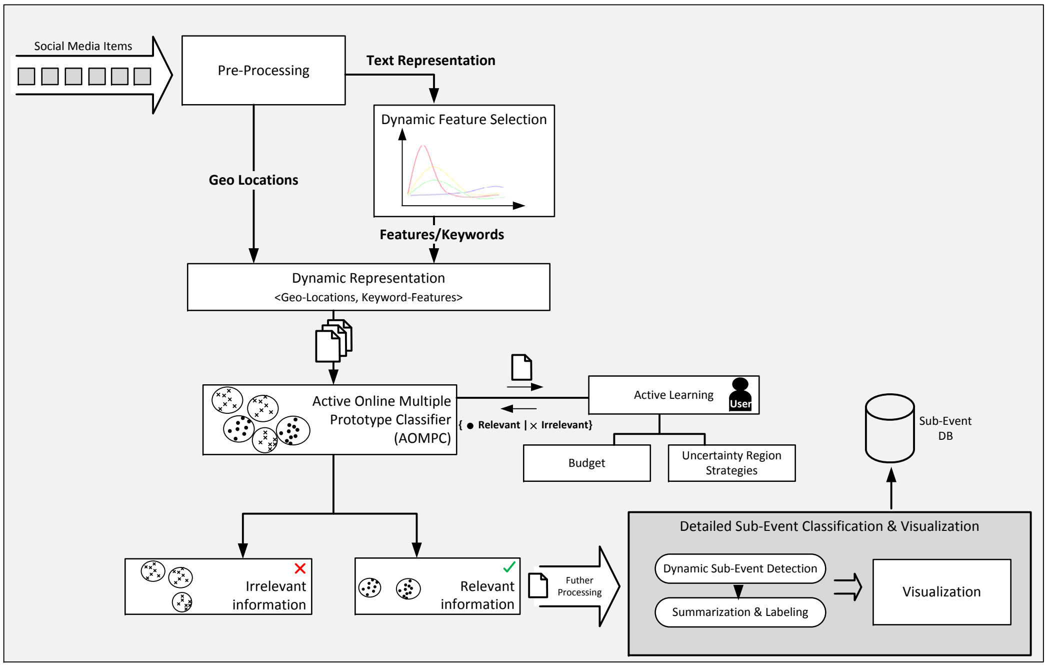

Social Media Items
Pre-Processing
Text Representation
Dynamic Feature Selection
Geo Locations
Features/Keywords
Dynamic Representation
<Geo-Locations, Keyword-Features>
Active Online Multiple Prototype Classifier (AOMPC)
{ ● Relevant | × Irrelevant}
Active Learning
User
Budget
Uncertainty Region Strategies
Sub-Event DB
Detailed Sub-Event Classification & Visualization
Dynamic Sub-Event Detection
Summarization & Labeling
Visualization
Irrelevant information
Relevant information
Futher Processing

Fig. 1. Processing steps

indicates where the message was issued in terms of coordinates, then we apply a combined distance measure $(dist\_text+dist\_geo)/2$. Specifically, $dist\_text$ refers to the Jaccard coefficient, while $dist\_geo$ is the Haversine distance [53], [5] described in Algorithm 2, steps 4-7. The coordinates are expressed in terms of latitude and longitude.

Moreover steps 4-12 of Algorithm 1 are related to the active learning part. The algorithm starts by checking whether the new input item lies in the uncertainty region between the relevant and irrelevant prototypes and whether there is enough budget for labeling this item. More details follow in the next section.

3.1 Definition of Budget

The idea of active learning is to ask for user feedback instead of labeling the incoming data item automatically. To limit the number of interventions of the user, a so called *budget*, is defined. Budget can be understood as the maximum number of queries to the user. We adapt the method presented in [63] to implement active learning in the context of online multiple prototype classification. In step 7 of Algorithm 1, the method *within_budget()* checks if enough budget is available for querying the user. The consumed budget after k items, b_k is defined in [63] as follows:

$$u_k = u_{k-1}\lambda + labeling_k;\ \lambda = (w-1)/w;\ b_k = \frac{u_k}{w} \tag{3}$$

where u_k estimates the amount of labels already queried by the system in the last w steps. The window w acts as memory [63] (e.g., last 100 item steps) described by λ. Hence, λ describes the fraction of including value u_{k-1}. $labeling_k$ updates u_k based on the requested label (i.e., $labeling_k = 0$ if no label was queried and $labeling_k = 1$ if there was a label

Algorithm 2 : $dist(\mathbf{v}, \mathbf{x})$

```
Input: Prototype v , input x
Output: Distance of (v,x)
 1: if the input is a social media item then
 2:     Compute the textual distance (Jaccard) as follows:
            $dist\_text = 1 - jaccard$, where:
            $jaccard = |A \cap B|/|A \cup B|$;
 3:     $distance = dist\_text$;
 4:     if the input is a composed social media item then
 5:         Compute the geo-location distance as follows:
                $dist\_geo = 1 - H(\mathbf{v}.geo\_co, \mathbf{x}.geo\_co)/\pi$
                where:
                $H(\mathbf{x}_1, \mathbf{x}_2) = 2 \cdot atan2(\sqrt{\phi}, \sqrt{1-\phi})$
                $\phi = sin^2(\frac{\Delta lat}{2}) + cos(\mathbf{x}_1.lat) \cdot cos(\mathbf{x}_2.lat) \cdot sin^2(\frac{\Delta lon}{2})$
                $\Delta lat = \mathbf{x}_2.lat - \mathbf{x}_1.lat$,
                $\Delta lon = \mathbf{x}_2.lon - \mathbf{x}_1.lon$
 6:         $distance = (dist\_geo + dist\_text)/2$;
 7:     end if
 8: else
 9:     Note: the input is no social media item
10:     Compute the Euclidean distance as follows:
            $dist\_Euclidean(\mathbf{v}, \mathbf{x}) = \sqrt{\sum_{i=1}^{M} (\mathbf{v}_i - \mathbf{x}_i)^2}$     (2)
11: end if
```

requested) for the current item k.

An upper bound B is defined describing the maximum number of requested labels. B is the fraction of data from window w that can be labeled (i.e., $B = 0.2$ are 20%). At each step, one input is processed. The *within_budget()* procedure in Algorithm 1 checks if enough budget is available (i.e., $b_k < B$). If so, the algorithm queries the label of the ambiguous input.

3.2 Which Data Items to Query?

In active learning, before querying the label, one has to decide which data points to query. Obviously one has to find those points, for which the classifier is not confident about the assignment decision (see Algorithm 1, step 7). In this paper, we use a simple mechanism based on the neighboring prototype proximity and labels. An input $\mathbf{x}$ is queried if its two most closest prototypes, $\mathbf{v}_i$ and $\mathbf{v}_j$ with distances φ_i and φ_j, respectively, and where $i = S(1)$ and $j = S(2)$, have different labels. Eq. 4 below formalizes the test which is called *simple conflicting neighborhood (SCN)* hereafter.

$$uncertainty(\mathbf{x}) = \begin{cases} 1 & \text{if } (|S| < 2) \text{ or} \\ & (|\varphi_i - \varphi_j| < UT \text{ and} \\ & \mathbf{v}_i.label \neq \mathbf{v}_j.label) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

However, to make the selection more constrained, a second variant is introduced. In fact, it is worthwhile to look at the border area of the inter-class uncertainty regions, where the labels are very important/useful. This border area could be used to track concept drift.

Eq. 5 shows the constraint by multiplying the threshold UT by a random number m that has a uniform distribution in unit interval [0,1] ($m \sim U(0,1)$) [63]. This variant is called *controlled variable conflicting neighborhood (CVCN).*

$$uncertainty(\mathbf{x}) = \begin{cases} 1 & \text{if } (|S| < 2) \text{ or} \\ & (|\varphi_i - \varphi_j| < (UT * m) \\ & \quad \text{and } \mathbf{v}_i.label \neq \mathbf{v}_j.label \\ & \quad \text{where } m \sim U(0,1)) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Moreover, the threshold UT can be continuously updated, as proposed in [63], according to the following rule:

$$\begin{cases} uncertainty(\mathbf{x}) = \begin{cases} 1 & \text{if } (|S| < 2) \text{ or} \\ & (|\varphi_i - \varphi_j| < UT \text{ and} \\ & \mathbf{v}_i.label \neq \mathbf{v}_j.label) \\ 0 & \text{otherwise} \end{cases} \\ UT = UT + (-1)^{uncertainty} * step \end{cases} \tag{6}$$

where *step* is set to 0.01 as suggested in [63]. We name this variant *dynamic conflicting neighborhood (DCN).* In the given equation it is combined with the *SCN* strategy. Additionally, we combined it with the *CVCN* strategy given above.

As a baseline for comparison, we implement a *random* version (see Eq. 7). We name this variant *random conflicting neighborhood (RCN).*

$$uncertainty(\mathbf{x}) = \begin{cases} 1 & \text{if } (|S| < 2) \text{ or} \\ & (|\varphi_i - \varphi_j| < r \\ & \quad \text{and } \mathbf{v}_i.label \neq \mathbf{v}_j.label \\ & \quad \text{where } r \sim U(0,1) \text{ is a} \\ & \quad \text{random variable)} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

We also implemented another version, called Random (R) that assumes a fixed uncertainty given by UT as shown in Eq. 8.

$$uncertainty(\mathbf{x}) = \begin{cases} 1 & \text{if } (|S| < 2) \text{ or} \\ & (r < UT) \\ & \quad \text{where } r \sim U(0,1) \text{ is a} \\ & \quad \text{random variable)} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

We ignore an absolute *pure random* version $r < B$, because it would increase the number of queries drastically compared to the other uncertainty variants.

3.3 Dynamic Representation of Social Media Stream

The SM items considered in our work are textual documents and therefore their representation will rely on the standard *tf-idf* [46], [37]. In this case, a document is represented as a bag-of-words. However, because social media documents arrive online and are processed as batches, *tf-idf* should be adapted to meet the streaming requirement [46]. Basically, the importance of a word is measured based on the number of incoming documents containing that word. Thus, the evolution of a term's importance should be reflected in the formulation of *tf-idf*. Here, we use a factor that scales *tf-idf* so that the importance increases and decreases according to the term's presence in the incoming batches (see Eq. 9).

$$scaled\_tf\_idf_{t,d} = importance_{t,\tau} \cdot tf_{t,d} \cdot idf_t \tag{9}$$

The importance factor $importance_{t,\tau}$ of term t is calculated over batches (windows) marked by time τ. The length of the batch is defined by the user (e.g., 30 minutes). It depends on the nature of the crisis. Slow evolution of the crisis may require longer windows, while fast evolution requires short windows. Terms with low importance value are removed from the index. For instance, if importance < 0.2, then 80% of the term's importance is lost. The importance of a term is computed as follows:

$$importance_{t,\tau} = g_{t,\tau}/g\_max_t \tag{10}$$

where $g_{t,\tau}$ is the weight of term t obtained at time τ. The weight $g_{t,\tau}$ is refreshed based on intermediate sampling intervals (i.e., sub-batches, like every 10 minutes). g_max_t is the maximum weight the term t reached. $g_{t,\tau}$ is expressed as follows:

$$g_{t,\tau} = \begin{cases} (1-\gamma) \cdot u_{t,\tau} + \gamma \cdot g_{t,\tau-1} & \text{if} \quad u_{t,\tau} > g_{t,\tau-1} \\ (1-\delta) \cdot u_{t,\tau} + \delta \cdot g_{t,\tau-1} & \text{otherwise} \end{cases} \tag{11}$$

where $u_{t,\tau}$ describes the incoming SM items containing t till time τ and $g_{t,\tau-1}$ is the weight of term t of the previous sampling interval $\tau - 1$. Case 1 of Eq. 11 shows how fast terms are learned (i.e., a smaller γ corresponds to faster increase of importance). Case 2 of Eq. 11 shows how fast terms should be forgotten (i.e., a higher δ corresponds to slower forgetting or decrease of importance). The values γ and δ are empirically set by the user. We suggest that $\gamma < \delta$ so that terms are learned faster, compared to forgetting them again.

4 Evaluation

In the following we present the experimental setting including the datasets and the metrics we used. We then describe the experiments and their outcomes.

4.1 Synthetic Datasets

To evaluate AOMPC, we use two synthetic datasets. The first one is a 2-dimensional numerical dataset and the second one is a collection of SM messages artificially generated by a tool. These datasets allow to observe the behavior of the algorithm, especially because it simulates data drift. The artificial SM data is used to evaluate the online classifier on geo-tagged textual data which is close to the real-world data.

The simple 2-dimensional synthetic dataset is based on Gaussian data (GD). GD consists of 4 batches (see Fig. 2) which are sequentially presented to AOMPC. Each batch consists of 200 points, generated by two Gaussians which actually represent two clusters. The upper clusters (100 points each), denoted as 'x', are assumed "irrelevant", while the lower clusters, denoted as 'o', are assumed "relevant". Batch-4 given in Fig. 2 contains a virtual or temporary drift caused by abrupt changes of the feature values [24].

The geo-tagged text collection, *synthetic social media dataset (SSMD)*, was generated using a tool[1] we originally developed for integrating SM into emergency exercises (i.e., training of first responders). We generated microblogs using a data generation tool we developed and which is based on a set of predefined text snippets that describe sub-events like "vehicles and garbage dumps on fire", "police attacked by rioters", and "shop on fire nearby" (see Fig. 3(a)). The randomly generated data follows the timeline of the UK riots (see [4]) described as an XML file (see Fig. 3(b)). This way we generate data which describes incidents close to what happened in reality. The XML file covers the different phases and particularly the sub-events of the UK riots which are marked as relevant or irrelevant using a tag (*relevant*) to provide the ground truth for the experiments. Irrelevant sub-events in the data are represented by real-world tweets collected from Twitter in relation to a given location (e.g., London), while relevant sub-events are based on the text snippets. On the other hand, additional data, in the form of textual annotations, was collected from Flickr and YouTube and was labeled based on the real-world sub-events of the riots (see [48]). In total, we used a collection of 1227 messages, mostly covering London districts. The data collected over 28 hours ('2011-08-06 19:44:00' to '2011-08-07 23:44:00') covers several calm periods during the riots. The data is split into 30-minutes batches to observe the behavior of AOMPC. The number of messages relevant to the riots is 312, with 116 distinct text messages. Furthermore, there are 915 irrelevant messages with 789 distinct messages. In all, the dataset contains approximately 322 repetitions of text messages. Repetition refers to messages that are very similar and correspond to retweets.

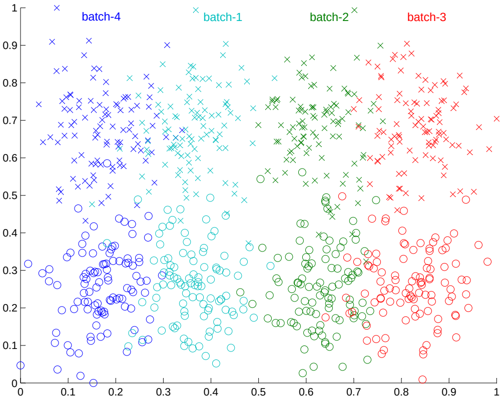


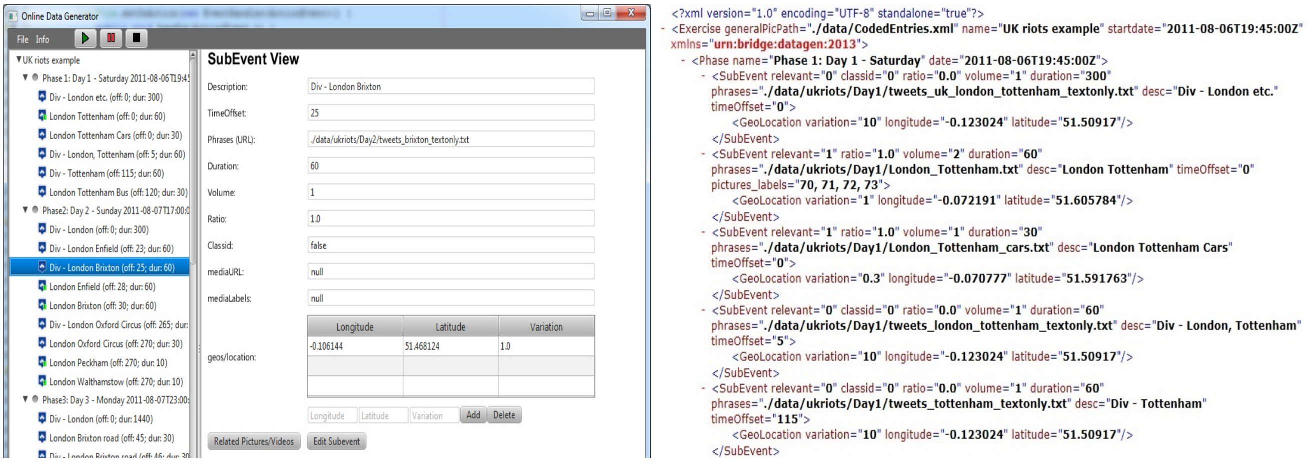
Fig. 2. GD dataset to simulate the stream appearing in the order batch-1, batch-2, batch-3, batch-4

4.2 Real-World Datasets

The CrisisLexT26 collection [41] was recently made available to the community. It consists of Twitter data related to 26 crises around the world. Each crisis is described by 1,000 items which were randomly selected and labeled through a crowdsourcing platform. The class labels of the items were assigned by the majority of three crowdsourcing workers. Four categories are available: *related to the crisis and informative*, *related to the crisis - but not informative*, *not related* and *not applicable*. In our case, we have considered items *relevant* only when they are labeled as *related to the crisis and informative*. Otherwise, they are considered irrelevant.

We selected two datasets from the CrisisLexT26 collection: Colorado Floods (CF) and Australia Bush-

[1] http://www.bridgeproject.eu/content/bridge_information_intelligence_flyer.pdf, [Accessed: August 2014]



UK riots example
- Phase 1: Day 1 - Saturday 2011-08-06T19:4
  - Div - London etc. (off: 0; dur: 300)
  - London Tottenham (off: 0; dur: 60)
  - London Tottenham Cars (off: 0; dur: 30)
  - Div - London, Tottenham (off: 5; dur: 60)
  - Div - Tottenham (off: 115; dur: 60)
  - London Tottenham Bus (off: 120; dur: 30)
- Phase2: Day 2 - Sunday 2011-08-07T17:00:0
  - Div - London (off: 0; dur: 300)
  - Div - London Enfield (off: 23; dur: 60)
  - Div - London Brixton (off: 25; dur: 60)
  - London Enfield (off: 28; dur: 60)
  - London Brixton (off: 30; dur: 60)
  - Div - London Oxford Circus (off: 265; dur:
  - London Oxford Circus (off: 270; dur: 30)
  - London Peckham (off: 270; dur: 10)
  - London Walthamstow (off: 270; dur: 10)
- Phase3: Day 3 - Monday 2011-08-07T23:00:
  - Div - London (off: 0; dur: 1440)
  - London Brixton road (off: 45; dur: 30)
  - Div - London Brixton road (off: 46; dur: 30

SubEvent View

Description: Div - London Brixton
TimeOffset: 25
Phrases (URL): ./data/ukriots/Day2/tweets_brixton_textonly.txt
Duration: 60
Volume: 1
Ratio: 1.0
Classid: false
mediaURL: null
mediaLabels: null
geos/location:

Longitude	Latitude	Variation
-0.106144	51.468124	1.0

(a) Data Generation Tool GUI

```
<?xml version="1.0" encoding="UTF-8" standalone="true"?>
- <Exercise generalPicPath="./data/CodedEntries.xml" name="UK riots example" startdate="2011-08-06T19:45:00Z"
  xmlns="urn:bridge:datagen:2013">
  - <Phase name="Phase 1: Day 1 - Saturday" date="2011-08-06T19:45:00Z">
    - <SubEvent relevant="0" classid="0" ratio="0.0" volume="1" duration="300"
      phrases="./data/ukriots/Day1/tweets_uk_london_tottenham_textonly.txt" desc="Div - London etc."
      timeOffset="0">
         <GeoLocation variation="10" longitude="-0.123024" latitude="51.50917"/>
      </SubEvent>
    - <SubEvent relevant="1" ratio="1.0" volume="2" duration="60"
      phrases="./data/ukriots/Day1/London_Tottenham.txt" desc="London Tottenham" timeOffset="0"
      pictures_labels="70, 71, 72, 73">
         <GeoLocation variation="1" longitude="-0.072191" latitude="51.605784"/>
      </SubEvent>
    - <SubEvent relevant="1" ratio="1.0" volume="1" duration="30"
      phrases="./data/ukriots/Day1/London_Tottenham_cars.txt" desc="London Tottenham Cars"
      timeOffset="0">
         <GeoLocation variation="0.3" longitude="-0.070777" latitude="51.591763"/>
      </SubEvent>
    - <SubEvent relevant="0" classid="0" ratio="0.0" volume="1" duration="60"
      phrases="./data/ukriots/Day1/tweets_london_tottenham_textonly.txt" desc="Div - London, Tottenham"
      timeOffset="5">
         <GeoLocation variation="10" longitude="-0.123024" latitude="51.50917"/>
      </SubEvent>
    - <SubEvent relevant="0" classid="0" ratio="0.0" volume="1" duration="60"
      phrases="./data/ukriots/Day1/tweets_tottenham_textonly.txt" desc="Div - Tottenham"
      timeOffset="115">
         <GeoLocation variation="10" longitude="-0.123024" latitude="51.50917"/>
      </SubEvent>
```

(b) UK riots stream in XML format

Fig. 3. Data Generation Tool

fires (AB) which are dated but not geo-tagged. CF data is from the period ’2013-09-12 07:00:00’ - ’2013-09-29 10:00:00’. The data is somewhat imbalanced, the number of relevant items is larger than that of the irrelevant ones. CF data consists of 751 relevant items and 224 irrelevant items and approximately 189 repetitions. Considering the number of relevant and irrelevant items of SSMD, CF has an opposite, but very similar, distribution. AB data is from the period ’2013-10-17 05:00:00’ - ’2013-10-29 12:30:00’. It consists of 645 relevant, 408 irrelevant items and approximately 385 retweets.

4.3 Evaluation Measures

Because AOMPC combines clustering and classification, we developed a combined performance measure, called *combined quality measure* (CQM), to evaluate the algorithms. It is defined as follows:

$$
\begin{aligned}
CQM = {} & [0.3 * \frac{\sum_{i=1}^{|Bt|} vm_i}{|Bt|}] + \\
& [0.5 * \frac{\sum_{i=1}^{|Bt|} (1 - er_i/100)}{|Bt|}] + \\
& [0.2 * (1 - (Q/\#items))]
\end{aligned}
\tag{12}
$$

It refers to two other known measures, namely the validity measure (VM) and the error-rate (ER) measure (see Appendix A for details). In terms of active learning budget B, the number of queries (Q) has been taken into account. In Eq. 12, Bt is the set of batches ($Bt = \{bt_1, \cdots, bt_{|Bt|}\}$) and vm_i and er_i are the values of VM and ER for batch bt_i respectively. *#items* is the number of items. As shown in Eq. 12, the measures are weighted based on their importance. ER is weighted with a factor of 0.5 due to its high importance, followed by VM with weight 0.3. Finally, the number of queries is weighted with 0.2. A high value of CQM corresponds to high clustering quality.

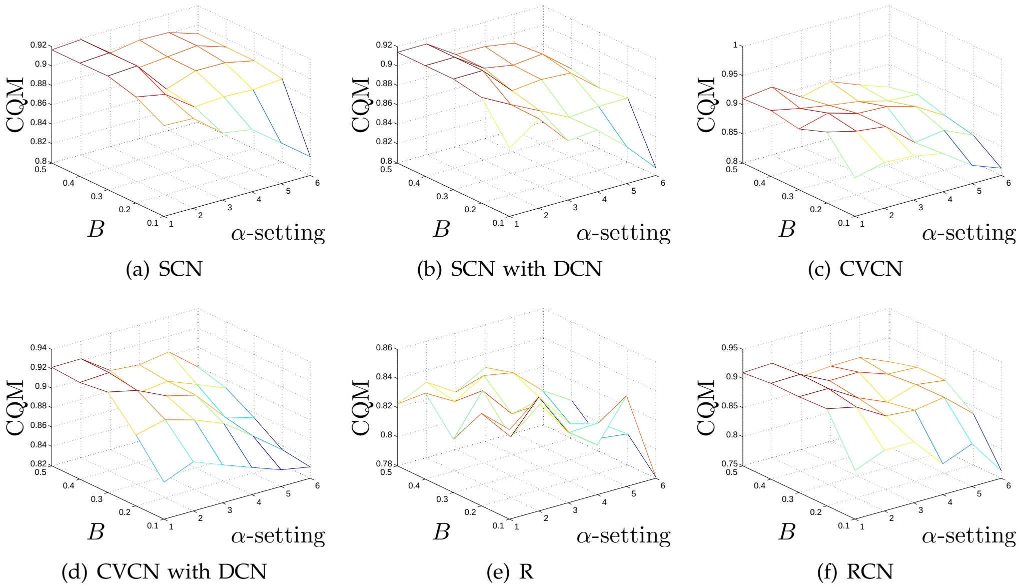
4.4 Experiments and Results

We conducted extensive analysis. In particular, we did a sensitivity analysis to observe the effect of the algorithm’s parameters: α, β, the threshold UT (see Alg. 1 and Tab. 1), and the budget B (see Sec. 3.1). In this section, we describe the outcome of the experiments on the datasets using different settings as shown in Tab. 2. We focus on the performance of the different uncertainty strategies using CQM. The *α-setting* represents the fixed and variable α settings.

TABLE 2
Evaluation Parameters

Parameter	Values/Instances
B	$B = 0.1, 0.2, \ldots 0.5$ with $w = 100$
UT	$0.1, 0.2, 0.3$
β	$1, 2, 3, 4$
fixed α	0.01 and 0.03
variable α	$\alpha = e^{\frac{-log(3)}{\beta}}$ as (1/3)-life-span $\alpha = e^{\frac{-log(2)}{\beta}}$ as (1/2)-life-span $\alpha = e^{\frac{log(2/3)}{\beta}}$ as (2/3)-life-span $\alpha = e^{\frac{log(7/8)}{\beta}}$ as (7/8)-life-span
Active Learning Method	SCN, CVCN, SCN with DCN, CVCN with DCN, R, and RCN
α-setting #1	equals to 0.01 (fixed α)
α-setting #2	equals to 0.03 (fixed α)
α-setting #3	equals to (1/3)-life-span (var. α)
α-setting #4	equals to (1/2)-life-span (var. α)
α-setting #5	equals to (2/3)-life-span (var. α)
α-setting #6	equals to (7/8)-life-span (var. α)

Gaussian Dataset (GD). Considering the most sensitive parameters, namely B and α (see Appendix B), the effect of active learning methods is illustrated in Fig. 4. The other parameters B and UT are discussed in Appendix B. In general it can be seen that the uncertainty strategy R yields the lowest CQM value and that RCN tends to query more often, since the pure random threshold r varies between 0 and 1 (see Sec. 3.2). For example, SCN has a query ratio of 0.14 and RCN a ratio of 0.2 to achieve a similar ER value (SCN with ER=1.250 and RCN with ER=1.370). On



(a) SCN (b) SCN with DCN (c) CVCN

(d) CVCN with DCN (e) R (f) RCN

Fig. 4. Results of the different active learning methods using the Gaussian data (GD) and the CQM measure.

average, SCN variants show the most stable results, while the CVCN variants slightly increase *CQM* for small values of B (i.e., $B \leq 0.2$), because they focus on concept drift near to the uncertainty boundary.

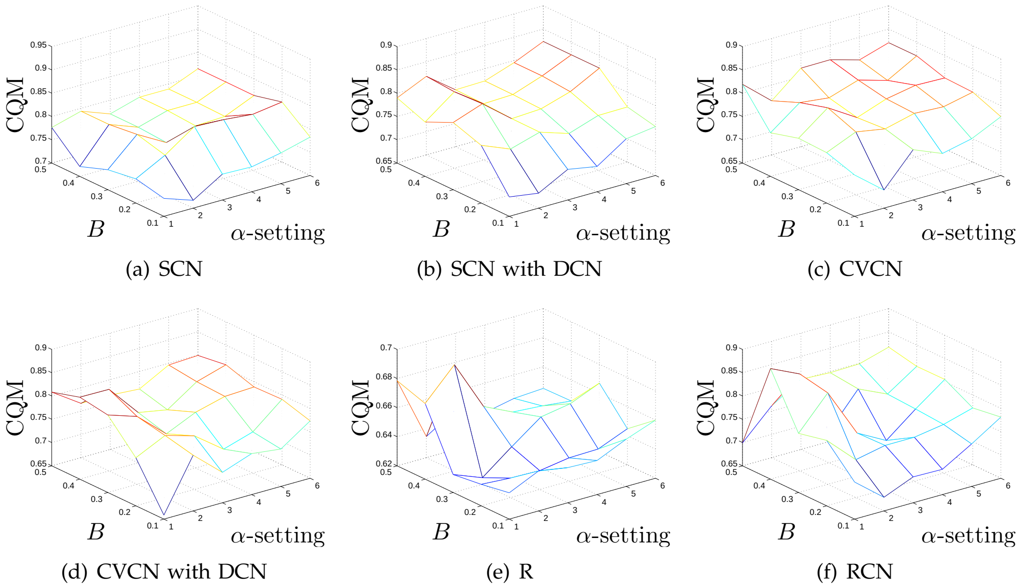
Synthetic Social Media Dataset (SSMD). The active learning strategies (SCN, CVCN, SCN with DCN and CVCN with DCN) given in Fig. 5 show that they outperform the random method R. Again, RCN shows good performance due to the higher variety of the threshold. For *CVCN with DCN* 0.22 queries and RCN 0.24 queries out of $B = 0.3$ are requested, reaching an ER of 7.3225 and 7.4984, respectively. A high value of B increases the overall quality of the results independently of the method (i.e., more labeled data is available to build the classification model). The *CVCN* options performs best for high values of B for the different α settings. In general, the active learning options *SCN with DCN* and *CVCN with DCN* perform best. This might indicate that concept drift appears along the uncertainty region border as those *"with DCN"* methods vary the border by changing UT. This behavior is expected, since data varies in a small range, i.e., geo-data within London area with similar incidents (damages caused by riots).

Colorado Floods (CF). Fig. 6 illustrates the outcome of AOMPC on the CF data for the different active learning strategies. The results of CF indicate good performance for the fixed α values and especially for a low budget B. The results corresponding to variable α are better than those obtained with fixed α. Note that higher α leads to fast update of the AOMPC prototypes and that variable α requires less queries (see Tab. 5). Based on the Levenshtein distance ($ldis$) ([33], for calculating similarity between character strings), there exist 105 items with similar text (i.e., $ldis \leq 0.2$) in CF, which is a quite small number. This also indicates that the length of the repeating text fragments are very small (105 vs. 189 repetitions of text). Therefore, the small number of similar items for this long period of the crisis and the performance related to the variable α with a fast adaptation are an indication that there are drifts in CF not near the inter-class border as defined by UT.

Australian Bushfires (AB). AOMPC's results on AB are illustrated in Fig. 7. The variable α shows nearly the same performance, but this time it is worse compared to the values obtained on CF. The AB dataset has a high amount of similar items, which is 582 (items with $ldis \leq 0.2$). This high amount of similar items is an indicator that changes in data are more common around the boundary, because similar vocabulary within the items is used. AOMPC shows the best performance with a fixed α value for all budget settings. Due to the high similarity between items combined with conflicting labels, it is more difficult to distinguish between relevant and irrelevant items. Consider the following example, which shows the same tweet, but labeled differently [41] (*Related-and-informative* and *Not-related*):

- Wed Oct 16 17:12:46 +0000 2013: "RT @Xxxxx: A dog has risked its life to save a litter of newborn kittens from a house fire in Melbourne, Australia http://t.co/Gz..",Eyewitness,Affected individuals,**Related and informative**
- Wed Oct 16 17:13:57 +0000 2013: "RT @Xxxxx: A dog has risked its life to save a litter of newborn kittens from a house fire in Melbourne, Australia http://t.co/Gz...",Not labeled,Not labeled,**Not related**

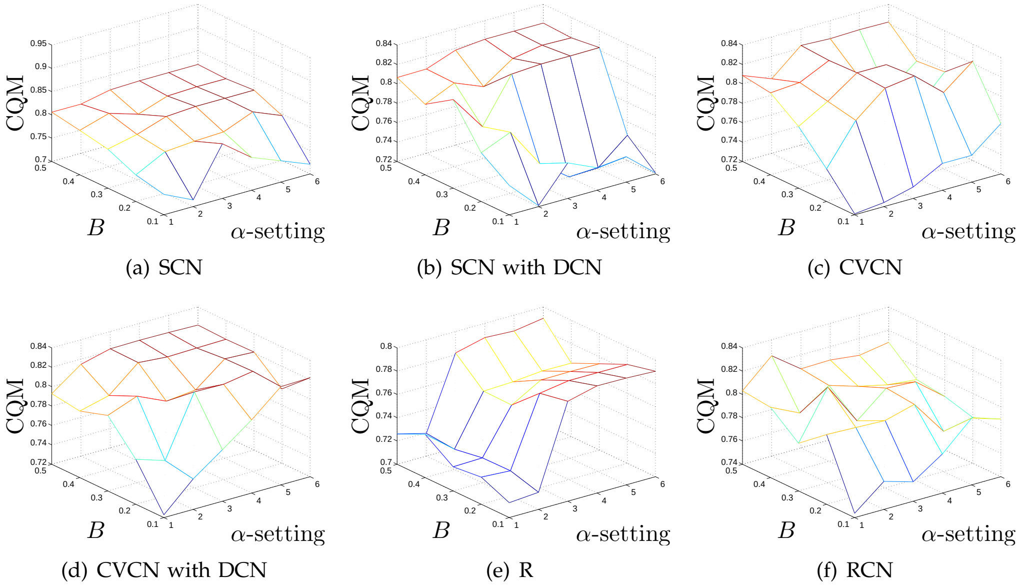
AB is an interesting dataset for testing the algorithms under various conditions. Fixed α provides much better quality on AB compared to other α-settings as shown in Fig. 7. Considering Figs. 7 and 6, we can conclude a fixed learning rate of α and "with DCN" active learning strategies produce good performance for both CF and AB, especially, for low values of B.



(a) SCN (b) SCN with DCN (c) CVCN

(d) CVCN with DCN (e) R (f) RCN

Fig. 5. Results of the different active learning methods using the synthetic social media dataset (SSMD) and the CQM measure



(a) SCN (b) SCN with DCN (c) CVCN

(d) CVCN with DCN (e) R (f) RCN

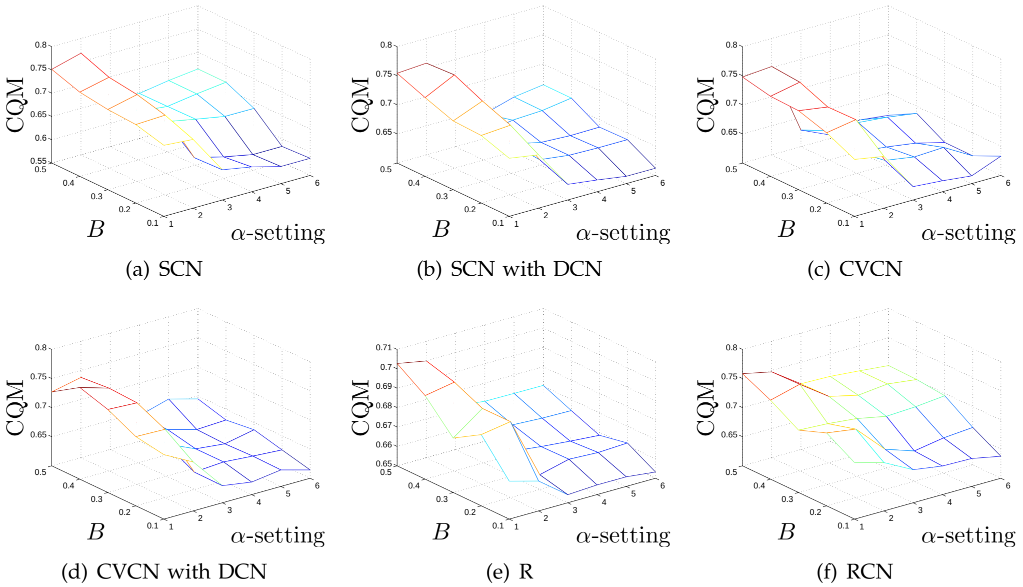
Fig. 6. Results of the different active learning methods using the Colorado Floods dataset (CF) and the CQM measure

4.5 Comparative Studies: AOMPC vs. Others

Beside the experiments with different datasets and parameters, we compare AOMPC against the unsupervised k-means algorithm that operates without labels and against a set of supervised online algorithms that require full labeling. This choice should help assess AOMPC against the extreme ends of the labeling spectrum:

- k-means: Given the online setting, the algorithm is run on batches of the data, setting the number of clusters to 10. For the real-world datasets (CF and AB) k-means has been initialized with 5 clusters, because there are fewer items per batch compared to the other datasets. For each batch $bt_i \in Bt$ of the data stream, the final centers obtained from the previous batch serve to initialize the centers of the current batch.
- Discriminative Online (Good?) Matlab Algorithms (DOGMA) [42]: The following algorithms are considered: PA-I [17], RBP and Perceptron [14], Projectron [44], Projectron++ [44], Forgetron (Kernel-Based Perceptron) [19], and Online Independent Support Vector Machines (OISVM) [43]. Because these algorithms are fully supervised, they are trained on all labeled data that is allowed by the budget B.

Running *k-means* on the different datasets produces the results shown in Tab. 3. CQM is calculated con-



(a) SCN (b) SCN with DCN (c) CVCN

(d) CVCN with DCN (e) R (f) RCN

Fig. 7. Results of the different active learning methods using the Australia Bushfires dataset (AB) and the CQM measure

TABLE 3
K-means: Avg. results for GD, SSMD, CF, and AB

	Q	VM	ER	CQM
GD	0	0.8270	2.8750	0.9337
SSMD	0	0.8143	4.7216	0.9207
CF	0	0.9608	0.9235	0.9836
AB	0	0.9477	1.3056	0.9778

sidering that k-means requires no queries ($Q = 0$). Items of a cluster are assigned the label of the majority. This assignment is performed after each batch and it is the base for computing the quality measures. It can be seen that for SSMD, k-means produces lower CQM compared to those of GD. This is also true in the case of AOMPC. Considering Fig. 4 and Fig. 5, it can be seen that AOMPC performs well. Comparing the results of k-means in Tab. 3 with the results of AOMPC in Tab. 5, the AOMPC values represent a good performance: AOMPC processes each data point only once and then discards it, whereas k-means uses all data points for computation. Clearly, the CQM values in Tab. 3 for CF and AB are very high, caused by low values of ER. For CF and AB, we used the same batch size (i.e., every 30 minutes) as for the generated SSMD dataset. More often, only a handful items are contained in the individual batches. Due to the small number of items per batch, it is not possible that relevant and irrelevant items are highly mixed within the created clusters of each batch. Hence, assignments are clear/unambigious.

TABLE 4
Best and ***worst*** CQM of DOGMA Algorithms (GD, SSMD, CF, AB)

		Q	B	VM	ER	CQM
GD	Forgetron	80	0.1	0.3029	32.5500	***0.6081***
	OISVM	80	0.1	0.8084	3.2625	**0.9062**
	RBP	160	0.2	0.3188	31.9500	***0.5959***
	OISVM	160	0.2	0.8217	2.9000	**0.8920**
	Forgetron	240	0.3	0.4100	25.3625	***0.6362***
	OISVM	240	0.3	0.8153	3.0250	**0.8695**
	RBP	320	0.4	0.2099	38.6750	***0.4896***
	OISVM	320	0.4	0.8180	2.9750	**0.8505**
	RBP	400	0.5	0.4811	20.9000	***0.6398***
	OISVM	400	0.5	0.8157	3.0250	**0.8296**
SSMD	PA-I	123	0.1	0.7228	5.4406	**0.8696**
	Projectron++	123	0.1	0.4202	11.5303	***0.7484***
	Projectron++	246	0.2	0.4105	10.5367	***0.7305***
	OISVM	246	0.2	0.8427	10.1921	**0.8619**
	PA-I	369	0.3	0.7636	2.2302	**0.8579**
	Forgetron	369	0.3	0.5593	9.7172	***0.7592***
	RBP	492	0.4	0.5025	9.0046	***0.7257***
	OISVM	492	0.4	0.8834	5.0767	**0.8596**
	PA-I	615	0.5	0.8647	1.2505	**0.8532**
	RBP	615	0.5	0.6244	5.3916	***0.7604***
CF	PA-I	98	0.1	0.7631	17.5100	**0.8214**
	Projectron++	98	0.1	0.7137	28.4213	***0.7520***
	PA-I	196	0.2	0.7728	15.9354	**0.8122**
	RBP	196	0.2	0.7141	23.7132	***0.7557***
	PA-I	294	0.3	0.8039	13.8672	**0.8118**
	Forgetron	294	0.3	0.7180	29.8722	***0.7060***
	PA-I	392	0.4	0.8222	12.7396	**0.8030**
	Forgetron	392	0.4	0.7117	28.5864	***0.6906***
	PA-I	490	0.5	0.8405	11.3371	**0.7955**
	Forgetron	490	0.5	0.7353	24.1613	***0.6998***
AB	PA-I	106	0.1	0.6791	22.9801	**0.7688**
	Projectron++	106	0.1	0.6440	32.6142	***0.7101***
	PA-I	212	0.2	0.7094	20.9924	**0.7678**
	Forgetron	212	0.2	0.6643	29.6821	***0.7109***
	PA-I	318	0.3	0.7428	17.6217	**0.7747**
	RBP	318	0.3	0.6707	27.3168	***0.7046***
	PA-I	424	0.4	0.7751	16.0927	**0.7721**
	Forgetron	424	0.4	0.6870	24.4803	***0.7037***
	Forgetron	530	0.5	0.7086	22.5930	***0.6996***
	OISVM	530	0.5	0.8087	13.6702	**0.7743**

The results of *DOGMA* algorithms related to the datasets are displayed in Tab. 4 for the best and worst cases. Details on the remaining algorithms can be found in Appendix C. Note that the DOGMA algorithms operate with the maximum amount of labels given by the budget. Hence, the training data is as large as the maximum number of items allowed by the budget. The CQM value is calculated such that $Q = B \cdot \#items$. The evaluation measures are computed based on each batch for comparison. DOGMA algorithms are trained based on randomly selected items from the dataset in advance. To ensure a fair comparison of DOGMA algorithms against AOMPC, we applied a 10-cross-validation strategy. The results in Tab. 4 show that in the case of GD, most of the DOGMA algorithms produce lower CQM compared to AOMPC results, which are illustrated in Fig. 4. It is an indication that the DOGMA algorithms are

TABLE 5
Best results of AOMPC based on budget B

	B	Query strategies	α (β for var. α)	Q (Q/#items)	VM	ER	CQM
GD	0.1	SCN	0.03	79.0 (0.10)	0.8460	2.3750	0.9222
	0.2	SCN	1/2 (4)	113.0 (0.14)	0.9180	1.2500	0.9409
	0.3	SCN	1/2 (4)	114.0 (0.14)	0.9180	1.2500	0.9406
	0.4	SCN	1/2 (4)	114.0 (0.14)	0.9180	1.2500	0.9406
	0.5	SCN	1/2 (4)	114.0 (0.14)	0.9180	1.2500	0.9406
SSMD	0.1	CVCN with DCN	0.03	113.0 (0.09)	0.7080	12.2120	0.8329
	0.2	SCN	1/3(1)	140.0 (0.11)	0.8440	12.2762	0.8690
	0.3	SCN	0.03	300.0 (0.24)	0.9161	8.8391	0.8817
	0.4	CVCN with DCN	0.01	256.0 (0.21)	0.8640	5.8791	0.8881
	0.5	CVCN with DCN	0.03	238.0 (0.19)	0.8876	9.4269	0.8804
CF	0.1	SCN	1/2 (2)	27.0 (0.03)	0.7451	18.0411	0.8278
	0.2	CVCN	1/2 (2)	32.0 (0.03)	0.7463	18.0141	0.8273
	0.3	**RCN**	2/3 (2)	223.0 (0.23)	0.8050	13.4949	0.8283
	0.4	SCN	0.03	297.0 (0.30)	0.8261	11.6488	0.8287
	0.5	SCN	0.03	297.0 (0.30)	0.8261	11.6488	0.8287
AB	0.1	CVCN with DCN	0.01	117.0 (0.11)	0.6669	31.4934	0.7204
	0.2	CVCN with DCN	0.03	215.0 (0.20)	0.7325	27.7243	0.7403
	0.3	SCN	0.01	304.0 (0.29)	0.7383	22.7398	0.7501
	0.4	CVCN with DCN	0.01	343.0 (0.33)	0.7607	18.8053	0.7690
	0.5	CVCN	0.03	380.0 (0.36)	0.7728	17.4619	0.7723

inefficient when dealing with changes in data, like the one artificially introduced in batch-4 of GD (see Fig. 2 of Sec. 4.1). In case of SSMD, CQM values obtained by most of the DOGMA algorithms (see Tab. 4) look similar to those values corresponding to the best active learning method of AOMPC (see Fig. 5 "with DCN" active learning methods). OISVM and PA-I produce the best performance on SSMD. In all, AOMPC performs well for on-the-fly querying. The DOGMA results related to CF and AB are also given in Tab. 4. Considering CQM as representative measure, DOGMA produced similar results to those produced by AOMPC shown in Figs. 6 and 7.

In a nutshell, AOMPC shows good performance compared to DOGMA, although the selection of items to query is performed on the fly. In addition, DOGMA algorithms use fully labeled data, while AOMPC uses only a subset of labeled data whose size is upper bounded by the budget.

4.6 Discussion and Future Work

The advantage of AOMPC compared to the other algorithms is the continuous processing of data streams and incremental update of knowledge, where the existing prototypes act as memory for the future. Here forgetting of outdated knowledge is controlled by α, which also depends on the budget. Learning serves to adapt and/or create clusters in a continuous way. The algorithm queries labels on-the-fly for continuously updating the classification model. In summary, it can be said that budget B and threshold UT are related to each other. Increasing their values increases the quality of the algorithm. B has also an influence on the number of clusters that are created (i.e., the more often the user is asked, the more hints for new clusters are given).

The advantage of our algorithm compared to the others is the transferred knowledge from one batch to the next creating a continuous view on the arriving data. The already known prototypes act as memory (i.e., forgetting is based on α and learning is based on the new creation of clusters, see Algorithm 1).

In terms of performance, Tab. 5 shows the best results of AOMPC for different budget values using the *CQM* measure. For GD, the variable learning rate α and the fixed α rate in the case of SSMD show good performance. For CF, the variable learning rate seems to be more suitable considering the number of queries. AOMPC produces good results on AB using a fixed learning rate. The reason is that the data items are very similar and that changes within the textual data happen slowly and near the boundary. Finally, comparing the active learning strategies (*"DCN"* options), we can notice that very good performance is achieved especially for SSMD and CF. The quality of clustering increases even for low values of B.

Overall, AOMPC shows a quite good performance (see Tables 4, 3 and 5), despite the fact that it operates online and handles labeling just-in-time. Moreover, AOMPC was run on batches just for the sake of feature selection (see Sec. 3.3). AOMPC can run in purely point-based online mode (i.e., item-by-item) as well. In the future, we plan to extend this algorithm by deleting clusters when they lose their importance. This could also be done for features in order to obtain an evolving feature space. We also plan to implement a variable budget strategy so that, for instance, the number of queries (i.e., budget) is bigger for cold-start and gets reduced afterward, depending on the uncertainty and the performance of the algorithm. Finally, it would be interesting to identify drift, without defining a threshold, but by considering the general case, where classes are non-contiguous.

5 Conclusion

This paper presents a streaming analysis framework for distinguishing between relevant and irrelevant data items. It integrates the user into the learning process by considering the active learning mechanism. We evaluated the framework for different datasets, with different parameters and active learning strategies. We considered synthetic datasets to understand the behavior of the algorithm and real-world social media datasets related to crises. We compared the proposed algorithm, AOMPC, against many existing algorithms to illustrate the good performance under different parameter settings. As explained in Sec. 4.6, the algorithm can be extended to overcome many issues, for instance by considering: dynamic budget, dynamic deletion of stale clusters, and generalization to handle non-contiguous class distribution.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n°261817 and was partly performed in the Lakeside Labs research cluster at Alpen-Adria-Universität Klagenfurt.

References

[1] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao, "Semantics + Filtering + Search = Twitcident. Exploring Information in Social Web Streams," in *Proc. of the 23rd ACM Conf. on Hypertext and Social Media*. ACM, 2012, pp. 285–294.

[2] U. Ahmad, A. Zahid, M. Shoaib, and A. AlAmri, "Harvis: An integrated social media content analysis framework for youtube platform," *Information Systems*, vol. 69, pp. 25 – 39, 2017.

[3] G. Backfried, J. Gollner, G. Qirchmayr, K. Rainer, G. Kienast, G. Thallinger, C. Schmidt, and A. Peer, "Integration of Media Sources for Situation Analysis in the Different Phases of Disaster Management: The QuOIMA Project," in *Eur. Intel. and Security Informatics Conf.*, Aug 2013, pp. 143–146.

[4] BBC News Europe. (2012, Aug.) England Riots: Maps and Timeline. [Online]. Available: http://www.bbc.co.uk/news/uk-14436499

[5] H. Becker, M. Naaman, and L. Gravano, "Learning Similarity Metrics for Event Identification in Social Media," in *Proc. of the Third ACM Int'l Conf. on Web Search and Data Mining*, ser. WSDM '10. NY, USA: ACM, 2010, pp. 291–300.

[6] J. Bezdek, T. Reichherzer, G. Lim, and Y. Attikiouzel, "Multiple-Prototype Classifier Design," *IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 28, no. 1, pp. 67–79, Feb 1998.

[7] M. Biehl, B. Hammer, and T. Villmann, "Prototype-based Models in Machine Learning," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 7, no. 2, pp. 92–111, 2016.

[8] A. Bouchachia, "Learning with Incrementality," in *Proc. of the Int'l Conf. on Neural Information Processing*, 2006, pp. 137–146.

[9] ——, "Incremental Learning with Multi-Level Adaptation," *Neurocomputing*, vol. 74, no. 11, pp. 1785–1799, 2011.

[10] A. Bouchachia and C. Vanaret, "Incremental Learning Based on Growing Gaussian Mixture Models," in *10th Int'l Conf. on Machine Learning and Applications and Workshops (ICMLA)*, vol. 2, Dec 2011, pp. 47–52.

[11] ——, "GT2FC: An Online Growing Interval Type-2 Self-Learning Fuzzy Classifier," *IEEE Transactions on Fuzzy Systems*, vol. 22, no. 4, pp. 999–1018, 2014.

[12] M.-R. Bouguelia, Y. Belaïd, and A. Belaïd, "An Adaptive Streaming Active Learning Strategy based on Instance Weighting," *Pattern Recognition Letters*, vol. 70, pp. 38 – 44, 2016.

[13] M. Büscher and M. Liegl, "Connected Communities in Crises," in *Social Media Analysis for Crisis Management*, H. Hellwagner, D. Pohl, and R. Kaiser, Eds. IEEE Computer Society Special Technical Community on Social Networking E-Letter, March 2014, vol. 2, no. 1.

[14] G. Cavallanti, N. Cesa-Bianchi, and C. Gentile, "Tracking the best Hyperplane with a simple Budget Perceptron," *Machine Learning*, vol. 69, no. 2-3, pp. 143–167, 2007.

[15] L. Chen, K. S. M. Tozammel Hossain, P. Butler, N. Ramakrishnan, and B. A. Prakash, "Syndromic Surveillance of Flu on Twitter Using Weakly Supervised Temporal Topic Models," *Data Mining and Knowledge Discovery*, vol. 0, no. 3, pp. 681–710, May 2016.

[16] T. M. Cover and J. A. Thomas, "Entropy, Relative Entropy and Mutual Information," in *Elements of Information Theory*. New Jersey: A John Wiley & Sons, 2006.

[17] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online Passive-Aggressive Algorithms," *J. Mach. Learn. Res.*, vol. 7, pp. 551–585, Dec. 2006.

[18] S. Dashti, L. Palen, M. P. Heris, K. M. Anderson, S. Anderson, and S. Anderson, "Supporting Disaster Reconnaissance with Social Media Data: A Design-Oriented Case Study of the 2013 Colorado Floods," in *Proc. of the 11th Int'l Conference on Information Systems for Crisis Response and Management*, University Park, Pennsylvania, USA, 2014.

[19] O. Dekel, S. Shalev-Shwartz, and Y. Singer, "The Forgetron: A Kernel-Based Perceptron on a Fixed Budget," in *NIPS*. MIT Press, 2005, pp. 259–266.

[20] A. Denecke, H. Wersing, J. Steil, and E. Körner, "Online figure-ground segmentation with adaptive metrics in generalized LVQ," *Neurocomputing*, vol. 72, no. 7-9, pp. 1470 – 1482, 2009.

[21] S. Denef, P. S. Bayerl, and N. Kaptein, "Social Media and the Police - Tweeting Practices of British Police Forces during the August 2011 Riots," in *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems (CHI)*, Paris, France, May 2013.

[22] N. Dufty, "Using Social Media to build Community Disaster Resilience," *The Australian Journal of Emergency Management*, vol. 27, no. 1, pp. 40–45, 2012.

[23] M. Freeman and A. Freeman, "Bonding over Bushfires: Social Networks in Action," in *IEEE International Symposium on Technology and Society (ISTAS)*, June 2010, pp. 419–426.

[24] J. a. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A Survey on Concept Drift Adaptation," *ACM Comput. Surv.*, vol. 46, no. 4, pp. 44:1–44:37, 2014.

[25] B. Hammer, D. Hofmann, F.-M. Schleif, and X. Zhu, "Learning Vector Quantization for (dis-)similarities," *Neurocomputing*, vol. 131, pp. 43 – 51, 2014.

[26] S. Hao, J. Lu, P. Zhao, C. Zhang, S. C. H. Hoi, and C. Miao, "Second-Order Online Active Learning and Its Applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 7, pp. 1338–1351, July 2018.

[27] S. Hao, P. Hu, P. Zhao, S. C. H. Hoi, and C. Miao, "Online Active Learning with Expert Advice," *ACM Trans. Knowl. Discov. Data*, vol. 12, no. 5, pp. 58:1–58:22, 2018.

[28] S. R. Hiltz, B. van de Walle, and M. Turoff, "The Domain of Emergency Management Information," in *Information Systems for Emergency Management*. Armonk, New York: B. van de Walle, M. Truoff and S. R. Hiltz, 2010, vol. 16, pp. 3–19.

[29] D. Ienco, A. Bifet, I. Žliobaitė, and B. Pfahringer, "Clustering Based Active Learning for Evolving Data Streams," in *Discovery Science*, ser. Lecture Notes in Computer Science, J. Fürnkranz, E. Hüllermeier, and T. Higuchi, Eds. Springer Berlin Heidelberg, 2013, vol. 8140, pp. 79–93.

[30] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, "AIDR: Artificial Intelligence for Disaster Response," in *Proc. of the Companion Publication of the 23rd Int'l Conf. on World Wide Web*, ser. WWW Companion '14, April 2014, pp. 159–162.

[31] Y. Ishikawa, Y. Chen, and H. Kitagawa, "An On-Line Document Clustering Method Based on Forgetting Factors," in *Research and Advanced Technology for Digital Libraries*, ser. Lecture Notes in Computer Science, P. Constantopoulos and I. T. Solvberg, Eds. Springer, 2001, vol. 2163, pp. 325–339.

[32] T. Kohonen, "The Self-Organizing Map," *Proc. of the IEEE*, vol. 78, no. 9, pp. 1464 –1480, Sep 1990.

[33] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," in *Soviet physics doklady*, vol. 10, 1966, p. 707.

[34] R. Li, K. H. Lei, R. Khadiwala, and K.-C. Chang, "TEDAS: A Twitter-based Event Detection and Analysis System," in *IEEE 28th Int'l Conf. on Data Engineering (ICDE)*, 2012, pp. 1273–1276.

[35] S. Liu, L. Palen, J. Sutton, A. Hughes, and S. Vieweg, "In Search of the Bigger Picture: The Emergent Role of On-Line Photo-Sharing in Times of Disaster," in *Proc. of the 5th Int'l ISCRAM Conf.*, 2008.

[36] L. Ma, S. Destercke, and Y. Wang, "Online Active Learning of Decision Trees with Evidential Data," *Pattern Recognition*, vol. 52, pp. 33 – 45, 2016.

[37] C. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[38] B. Mokbel, B. Paassen, F.-M. Schleif, and B. Hammer, "Metric Learning for Sequences in Relational LVQ," *Neurocomputing*, vol. 169, pp. 306 – 322, 2015.

[39] B. Mozafari, P. Sarkar, M. Franklin, M. Jordan, and S. Madden, "Scaling Up Crowd-sourcing to Very Large Datasets: A Case for Active Learning," *Proc. VLDB Endow.*, vol. 8, no. 2, pp. 125–136, Oct. 2014.

[40] V. K. Neppalli, C. Caragea, A. Squicciarini, A. Tapia, and S. Stehle, "Sentiment Analysis during Hurricane Sandy in Emergency Response," *International Journal of Disaster Risk Reduction*, vol. 21, pp. 213 – 222, 2017.

[41] A. Olteanu, S. Vieweg, and C. Castillo, "What to Expect When the Unexpected Happens: Social Media Communications Across Crises," *In Proc. of the ACM Conf. on Computer Supported Cooperative Work and Social Computing*, 2015.

[42] F. Orabona, *DOGMA: a MATLAB toolbox for Online Learning*, 2009, software available at http://dogma.sourceforge.net.

[43] F. Orabona, C. Castellini, B. Caputo, L. Jie, and G. Sandini, "On-line Independent Support Vector Machines," *Pattern Recognition*, vol. 43, no. 4, pp. 1402 – 1412, 2010.

[44] F. Orabona, J. Keshet, and B. Caputo, "Bounded Kernel-Based Online Learning," *J. Mach. Learn. Res.*, vol. 10, pp. 2643–2666, Dec. 2009.

[45] S.-Y. Perng, M. Büscher, L. Wood, R. Halvorsrud, M. Stiso, L. Ramirez, and A. Al-Akka, "Peripheral Response: Microblogging During the 22/7/2011 Norway Attacks," *Int'l Journal of Information Systems for Crisis Response and Management (IJISCRAM)*, vol. 5, no. 1, pp. 41–57, 2013.

[46] D. Pohl, A. Bouchachia, and H. Hellwagner, "Online Processing of Social Media Data for Emergency Management," in *Int'l Conf. on Machine Learning and Applications*, vol. 2, Dec. 2013, pp. 333 – 338.

[47] D. Pohl, "Social Media Analysis for Crisis Management: A Brief Survey," in *Social Media Analysis for Crisis Management*, H. Hellwagner, D. Pohl, and R. Kaiser, Eds. IEEE Computer Society Special Technical Community on Social Networking E-Letter, March 2014, vol. 2, no. 1.

[48] D. Pohl, A. Bouchachia, and H. Hellwagner, "Social Media for Crisis Management: Clustering Approaches for Sub-Event Detection," *Multimedia Tools and Applications*, pp. 1–32, 2013.

[49] ——, "Online Indexing and Clustering of Social Media Data for Emergency Management," *Neurocomputing*, vol. 172, pp. 168 – 179, 2016.

[50] J. R. Ragini, P. R. Anand, and V. Bhaskar, "Mining Crisis Information: A Strategic Approach for Detection of People at Risk through Social Media Analysis," *International Journal of Disaster Risk Reduction*, vol. 27, pp. 556 – 566, 2018.

[51] C. Reuter and M. Kaufhold, "Fifteen Years of Social Media in Emergencies: A Retrospective Review and Future Directions for Crisis Informatics," *Journal of Contingencies and Crisis Management*, vol. 26, no. 1, pp. 41–57, 2018.

[52] T. Reuter and P. Cimiano, "Event-based Classification of Social Media Streams," in *Proc. of the 2nd ACM Int'l Conf. on Multimedia Retrieval*, 2012, pp. 22:1–22:8.

[53] T. Reuter, P. Cimiano, L. Drumond, K. Buza, and L. Schmidt-Thieme, "Scalable Event-Based Clustering of Social Media via Record Linkage Techniques," in *The 5th Int'l Conf. on Weblogs and Social Media*, 2011, pp. 313–320.

[54] A. Rosenberg and J. Hirschberg, "V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure," in *EMNLP-CoNLL*, vol. 7, 2007, pp. 410–420.

[55] B. Settles, "Active Learning Literature Survey," *University of Wisconsin, Madison*, vol. 52, pp. 55–66, 2010.

[56] E. Shook, K. Leetaru, G. Cao, A. Padmanabhan, and S. Wang, "Happy or Not: Generating Topic-based Emotional Heatmaps for Culturomics using CyberGIS," in *2012 IEEE 8th Int'l Conf. on E-Science*, oct. 2012, pp. 1 –6.

[57] J. Smailović, M. Grčar, N. Lavrač, and M. Žnidaršič, "Stream-based Active Learning for Sentiment Analysis in the Financial Domain (in press)," *Information Sciences*, April 2014.

[58] K. Starbird and J. Stamberger, "Tweak the Tweet: Leveraging Microblogging Proliferation with a Prescriptive Syntax to Support Citizen Reporting," in *Proc. of the 7th Int'l ISCRAM Conf.*, Seattle, USA, May 2010.

[59] T. Terpstra, A. de Vries, R. Stronkman, and G. L. Paradies, "Towards a Realtime Twitter Analysis during Crises for Operational Crisis Management," in *Proc. of the 9th Int'l ISCRAM Conf.*, Vancouver, April 2012.

[60] M. F. Umer and M. S. H. Khiyal, "Classification of Textual Documents using Learning Vector Quantization," *Information Technology Journal*, vol. 6, no. 1, pp. 154–159, 2007.

[61] S. Vieweg and A. Hodges, "Rethinking Context: Leveraging Human and Machine Computation in Disaster Response," *Computer*, vol. 47, no. 4, pp. 22–27, Apr 2014.

[62] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness," in *Proc. of the Int'l Conf. on Human Factors in Computing Systems*, ser. CHI '10. NY, USA: ACM, 2010, pp. 1079–1088.

[63] I. Žliobaitė, A. Bifet, B. Pfahringer, and G. Holmes, "Active Learning With Drifting Streaming Data," *IEEE Trans. on Neural Networks and Learning Sys.*, vol. 25, no. 1, pp. 27–39, Jan 2014.

[64] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, 2011.

[65] J. Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power, "Emergency Situation Awareness from Twitter for Crisis Management," *Int'l Workshop on Social Web for Disaster Management (SWDM), In conjunction with WWW'12*, no. 99, p. 1, 2012.

[66] ——, "Using Social Media to Enhance Emergency Situation Awareness," *IEEE Intelligent Sys.*, vol. 27, no. 6, pp. 52–59, 2012.

Daniela Pohl received her Dipl.-Ing. (Master's degree) in Computer Science in 2008 at the Alpen-Adria-Universität Klagenfurt, Austria. She worked as research assistant in the scope of the EU-funded FP7 project BRIDGE (www.bridgeproject.eu) to develop technical solution to improve crisis management. She received her doctoral degree 2015 at the Alpen-Adria-Universität Klagenfurt. Her research interests include information retrieval and machine learning.

Abdelhamid Bouchachia is currently an Associate Professor at Bournemouth University, Department of Computing, Smart Technology Research Centre, UK. His major research interests include Machine Learning and Soft Computing with a particular focus on online/incremental learning, semi-supervised learning, prediction systems, and uncertainty modeling. He is the general chair of the International Conference on Adaptive and Intelligent Systems (ICAIS). He serves as program committee member for many conferences. He also serves as Associate Editor of Evolving Systems and acts as member of Evolving Intelligent Systems (EIS) Technical Committee (TC) of the IEEE Systems, Man and Cybernetics Society, the IEEE Task-Force for Adaptive and Evolving Fuzzy Systems and the IEEE Computational Intelligence Society.

Hermann Hellwagner is a full professor of Informatics in the Institute of Information Technology (ITEC), Klagenfurt University, Austria, leading the Multimedia Communications group. His current research areas are distributed multimedia systems, multimedia communications, and quality of service. He has received many research grants from national (Austria, Germany) and European funding agencies as well as from industry, is the editor of several books, and has published more than 250 scientific papers on parallel computer architecture, parallel programming, and multimedia communications and adaptation. He is a senior member of the IEEE, member of the ACM and OCG (Austrian Computer Society); he was Vice President of the Austrian Science Fund (FWF).

Appendix A Details on Evaluation Measures

For the evaluation, the purity (P) and entropy (E) measures have been computed. P and E are defined as follows [48]:

$$P_i = \frac{1}{n_i} Max_j(n_{i,j}) \tag{13}$$

$$E_i = -\frac{1}{log(H)} \sum_{j=1}^{H} \frac{n_{i,j}}{n_i} log \frac{n_{i,j}}{n_i} \tag{14}$$

where n_i is the total number of data items in cluster i and $n_{i,j}$ is the number of items of class j in cluster i. High *purity* values indicate good clustering quality (i.e., correct assignment of classes to clusters). In contrast, small *entropy* values indicate high quality in clustering. We used the average P and E value for all identified clusters m as follows:

$$P = \frac{1}{m} \sum_{i=1}^{m} P_i \tag{15}$$

$$E = \frac{1}{m} \sum_{i=1}^{m} E_i \tag{16}$$

The *validity-measure* (VM) [54] is the harmonic mean of two quantities: *homogeneity* and *completeness*. VM is defined as follows:

$$VM = \frac{(1+\kappa) * homogeneity * completeness}{(\kappa * homogeneity) + completeness} \tag{17}$$

where κ is a weighting parameter. The homogeneity is given by:

$$homogeneity = \begin{cases} 1 & \text{if } H(C|K) = 0 \\ 1 - \frac{H(C|K)}{H(C)} & \text{otherwise} \end{cases} \tag{18}$$

such that:

$$H(C|K) = -\sum_{k=1}^{|K|} \sum_{c=1}^{|C|} \frac{a_{c,k}}{N} log \frac{a_{c,k}}{\sum_{c=1}^{|C|} a_{c,k}}$$

$$H(C) = \sum_{c=1}^{|C|} \frac{\sum_{k=1}^{|K|} a_{c,k}}{N} log \frac{\sum_{k=1}^{|K|} a_{c,k}}{N}$$

where N is the size of the dataset, $C = \{c_1, \dots, c_n\}$ is the set of n classes, $K = \{k_1, \dots, k_m\}$ is the set of m clusters. $A = a_{i,j}$ is the corresponding contingency table, where $a_{i,j}$ is the number of data points related to class c_i and cluster k_j. The perfect homogeneity (i.e., each cluster contains items of one single class) is reached when the conditional entropy $H(C|K) = 0$ [54]. $H(C|K)$ is maximum when the clustering does not give any new information (= $H(C)$ entropy). If there is only one class, $H(C) = 0$ and the homogeneity is 1.

Completeness is defined as follows:

$$completeness = \begin{cases} 1 & \text{if } H(K|C) = 0 \\ 1 - \frac{H(K|C)}{H(K)} & \text{otherwise} \end{cases} \tag{19}$$

where

$$H(K|C) = -\sum_{c=1}^{|C|} \sum_{k=1}^{|K|} \frac{a_{c,k}}{N} log \frac{a_{c,k}}{\sum_{k=1}^{|K|} a_{c,k}}$$

$$H(K) = \sum_{k=1}^{|K|} \frac{\sum_{c=1}^{|C|} a_{c,k}}{N} log \frac{\sum_{c=1}^{|C|} a_{c,k}}{N}$$

Completeness is expressed by the cluster assignments of the items in a class [54]. Based on the conditional entropy $H(K|C)$, a perfect completeness (i.e., a class is represented only by one cluster) is reached when the conditional entropy is 0. In contrast, when each class is split into all clusters then $H(K|C) = H(K)$, i.e., completeness is 0. In the special case of one cluster ($H(K) = 0$), the completeness is 1.

We are interested more in the homogeneity of the validity measure and therefore we set κ in Eq. 17 to 0.001 [54].

On the other hand, *Normalized Mutual Information* (NMI) quantifies the consistency between two clusterings $A = \{a_1, a_2, \dots\}$ and $C = \{c_1, c_2, \dots\}$. It is given as follows [48], [16]:

$$NMI(A;C) = \frac{MI(A;C)}{\sqrt{[H(A)H(C)]}} \tag{20}$$

where:

$$MI(A;C) = \sum_i \sum_j \frac{|a_i \cap c_j|}{N} log \frac{N|a_i \cap c_j|}{|a_i||c_j|} \tag{21}$$

$$MI(A;C) = H(A) + H(C) - H(A;C)$$

$$H(A) = -\sum_i \frac{|a_i|}{N} log \frac{|a_i|}{N}$$

The higher the NMI is, the more similar are the clusterings.

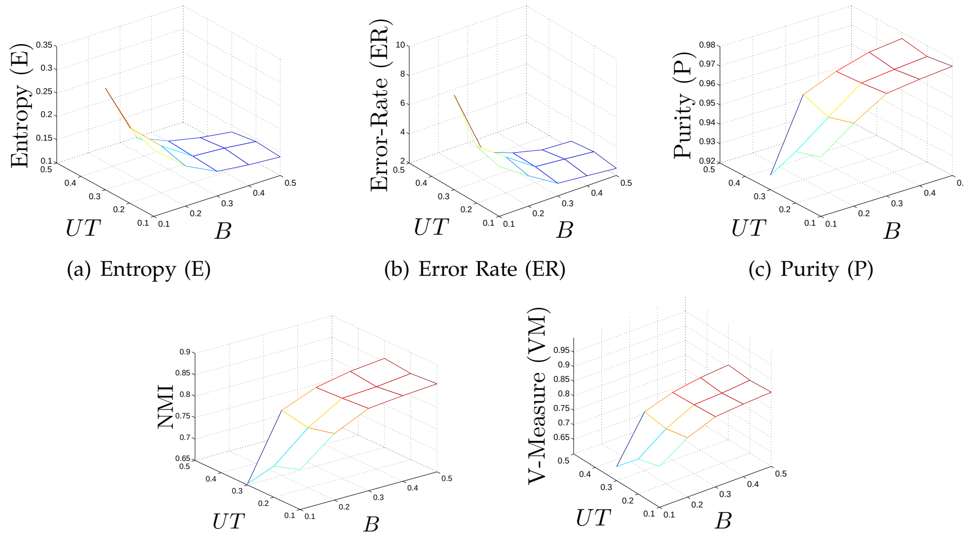
The error-rate (ER) described by Eq. 22 is the proportion (in percent) of incorrectly classified data items. If $L = \{l_1, l_2, \dots, l_N\}$ is the list of labels predicted by the classifier and $G = \{g_1, g_2, \dots, g_N\}$ is the list of true labels for N items, then:

$$ER(L;G) = \frac{|L \neq G|}{N} * 100 \tag{22}$$

The initial experiments (see Appendix B) showed that VM and ER are the most indicative measures that summarize - to a large extent - the other metrics. They are good indicators for the classification quality (e.g., see Fig. 8 for GD and Fig. 11 for SSMD).

Appendix B Additional Results

Gaussian Dataset (GD). Fig. 8 shows the effect of various combinations of the parameters UT and B on



(a) Entropy (E)

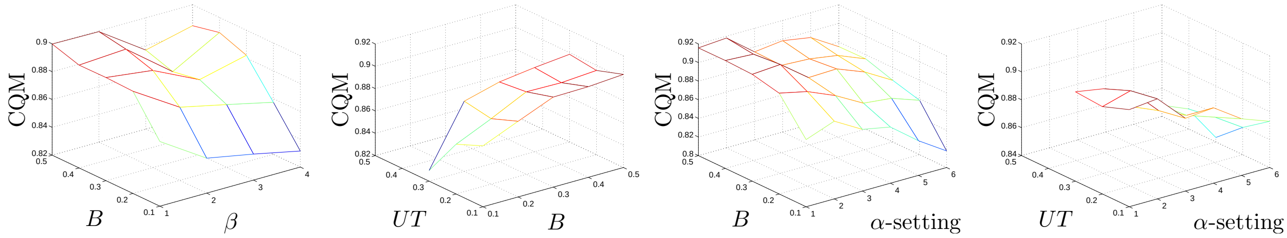
(b) Error Rate (ER)

(c) Purity (P)

(d) NMI

(e) Validity-measure (VM)

Fig. 8. Effect of UT and B on the accuracy using the Gaussian data (GD)



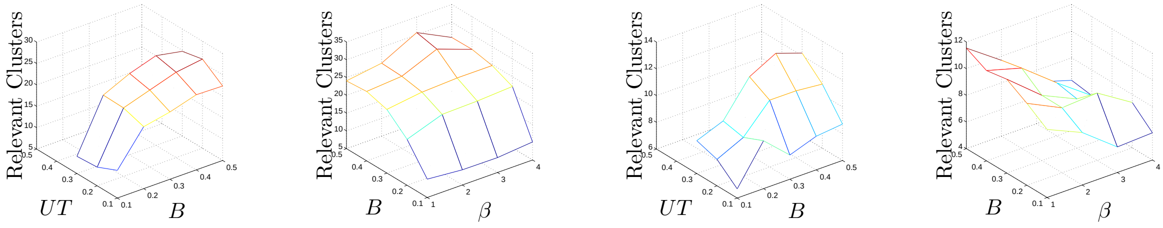
(a) B vs. β

(b) UT vs. B

(c) B vs. α-setting

(d) UT vs. α-setting

Fig. 9. CQM for various parameters using the Gaussian data (GD)



(a) GD: UT vs. B

(b) GD: B vs. β

(c) SSMD: UT vs. B

(d) SSMD: B vs. β

Fig. 10. Effect of parameters on the number of relevant clusters using the Gaussian data (GD) and the synthetic social media dataset (SSMD)

the clustering quality expressed by entropy (E), error-rate (ER), purity (P), Normalized Mutual Information (NMI) and validity-measure (VM).

It can be seen from Figs. 8(a) and 8(b) that E and ER decrease as B grows (i.e., more labeled data improves the classification). $B > 0.3$ does not significantly improve the results further. Hence, querying effort can be saved. Considering the NMI and VM measures in Figs. 8(d), and 8(e) for $B \geq 0.3$,

In Fig. 9(a), we compare β (for variable α) with the the budget B. Increasing β does not affect the quality of the classification much; hence, low values for β (i.e., 1-2) are more suitable for this dataset, but also $\beta = 4$ shows rather good performance. Fig. 13(a) shows β in relation to variable learning rates α. $\beta = 1$ to 2 performs better for most variable α rates, but $\beta = 4$ also yields good results when using smaller α. This behavior can be explained by the fact that only 4 batches/periods exist and an explicit drift happens in last batch of GD (see Fig. 2 in Section 4.1).

As shown in Fig. 10, the number of relevant clusters generated depends also on B and UT, while β seems to have less effect. In particular for $B > 0.3$, the number of relevant clusters tends to be stable and higher values of UT tend to generate more clusters. Moreover, Fig. 9(b) which is related to the combined quality measure (CQM) also indicates that $B > 0.3$ does not improve the results significantly.

We also compared different settings for the learning rate α (see Tab. 2) in relation to B and UT. Figs. 9(c) and 9(d) show that the fixed learning rates (0.01 and 0.03) yield slightly better results compared to the variable α-setting when the budget B is low. In addition, the higher UT, the better are the results. Following this set of experiments, we can state that the budget B and α are the most important parameters.

Synthetic Social Media Dataset (SSMD). Fig. 11 shows the clustering quality measures for different parameters UT and B. In Fig. 11, it can be seen that the higher the budget B and the threshold UT, the better are the results. Especially, ER is reduced (see Fig. 11(b)) and NMI and VM are increased as Figs. 11(d) and 11(e) show. This is also valid for CQM in Fig. 12(b).

Comparing β with B in Fig. 12(a), the same is true: the higher B, the better the results. Increasing β and B shows a slight positive effect on the quality of the overall results. Fig. 13(b) shows a slight increase of CQM values for higher values of β. Also, B and UT influence the number of relevant clusters produced (see Fig. 10). Relevant clusters indicate those clusters that consist of relevant tweets. As in the case of GD, the amount of relevant clusters depends mainly on the budget B (see Fig. 10).

More interesting is the tradeoff between α and B given in Fig. 12(c) and between α and UT given in Fig. 12(d). For this synthetic dataset, when α is set to 1 and to 2, a more positive effect on the results is obtained. This can also be observed for different active learning strategies.

APPENDIX C
COMPARATIVE STUDY: DOGMA DETAILS

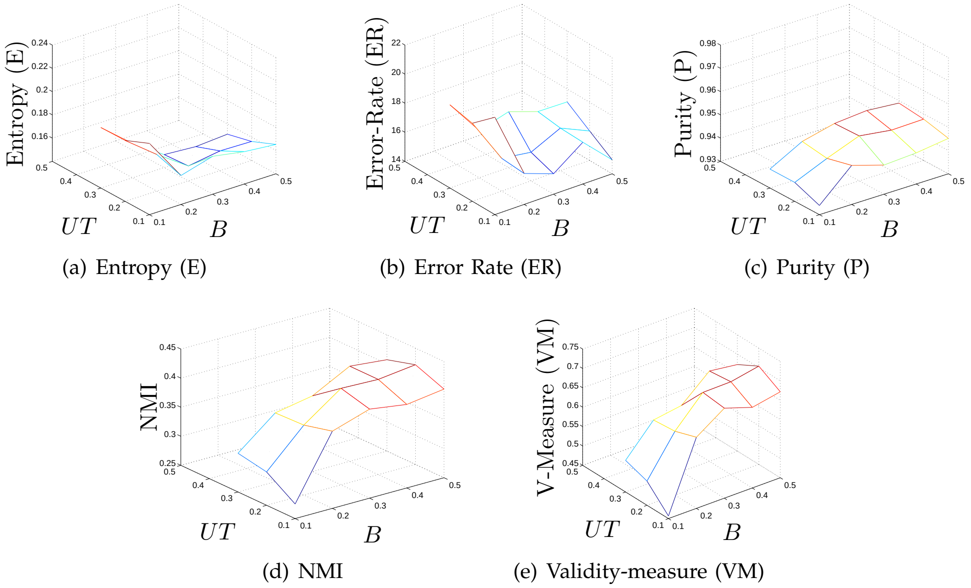
Tables 6, 7, 8, and 9 show the results related to the DOGMA algorithms, obtained on the different datasets: GD, SSMD, CF, and AB.

TABLE 6
DOGMA Algorithms (GD): Avg. results marking the **best** and ***worst*** CQM

	P	E	NMI	VM	ER	CQM
		DOGMA ($B = 0.1$)				
PA-I	0.9457	0.2530	0.7230	0.7269	6.1500	0.8673
Perceptron	0.9213	0.2798	0.6427	0.6577	11.4625	0.8200
Projectron	0.9153	0.2974	0.6198	0.6358	12.2375	0.8096
Projectron++	0.9370	0.2743	0.6915	0.6973	7.4000	0.8522
RBP	0.7920	0.5012	0.3517	0.3800	28.0250	0.6539
Forgetron	0.7453	0.5813	0.2745	0.3029	32.5500	***0.6081***
OISVM	0.9682	0.1886	0.8079	0.8084	3.2625	**0.9062**
		DOGMA ($B = 0.2$)				
PA-I	0.9615	0.2133	0.7802	0.7811	4.0125	0.8743
Perceptron	0.9527	0.2356	0.7482	0.7509	5.2250	0.8591
Projectron	0.9536	0.2348	0.7505	0.7529	5.0750	0.8605
Projectron++	0.9598	0.2186	0.7736	0.7748	4.2125	0.8714
RBP	0.7443	0.5764	0.2935	0.3188	31.9500	***0.5959***
Forgetron	0.7516	0.5814	0.3372	0.3529	28.6375	0.6227
OISVM	0.9711	0.1778	0.8216	0.8217	2.9000	**0.8920**
		DOGMA ($B = 0.3$)				
PA-I	0.9614	0.2106	0.7816	0.7829	4.0750	0.8545
Perceptron	0.9206	0.2901	0.6373	0.6514	11.2375	0.7792
Projectron	0.9216	0.2867	0.6409	0.6550	11.1375	0.7808
Projectron++	0.9594	0.2152	0.7745	0.7761	4.3625	0.8510
RBP	0.8280	0.4432	0.5009	0.5115	19.7500	0.6947
Forgetron	0.7642	0.5571	0.4021	0.4100	25.3625	***0.6362***
OISVM	0.9699	0.1840	0.8152	0.8153	3.0250	**0.8695**
		DOGMA ($B = 0.4$)				
PA-I	0.9599	0.2157	0.7754	0.7767	4.2375	0.8318
Perceptron	0.9482	0.2390	0.7355	0.7399	6.0375	0.8118
Projectron	0.9529	0.2310	0.7511	0.7540	5.2375	0.8200
Projectron++	0.9585	0.2217	0.7690	0.7704	4.3875	0.8292
RBP	0.6829	0.6790	0.1826	0.2099	38.6750	**0.4896**
Forgetron	0.7745	0.5217	0.3403	0.3676	29.3375	0.5836
OISVM	0.9704	0.1815	0.8180	0.8180	2.9750	**0.8505**
		DOGMA ($B = 0.5$)				
PA-I	0.9561	0.2251	0.7619	0.7639	4.7375	0.8055
Perceptron	0.9494	0.2306	0.7413	0.7466	6.1875	0.7930
Projectron	0.9492	0.2306	0.7410	0.7463	6.2375	0.7927
Projectron++	0.9541	0.2306	0.7544	0.7568	5.0000	0.8020
RBP	0.8408	0.4331	0.4603	0.4811	20.9000	***0.6398***
Forgetron	0.8730	0.3850	0.5502	0.5622	15.4250	0.6915
OISVM	0.9699	0.1838	0.8157	0.8157	3.0250	**0.8296**

TABLE 7
DOGMA Algorithms (SSMD): Avg. results marking the **best** and ***worst*** CQM

	P	E	NMI	VM	ER	CQM
		DOGMA ($B = 0.1$)				
PA-I	0.9668	0.1210	0.4349	0.7228	5.4406	**0.8696**
Perceptron	0.9550	0.1502	0.2763	0.4234	10.6493	0.7538
Projectron	0.9550	0.1502	0.2763	0.4234	10.6554	0.7537
Projectron++	0.9538	0.1524	0.2735	0.4202	11.5303	***0.7484***
RBP	0.9535	0.1534	0.2843	0.4518	11.2847	0.7591
Forgetron	0.9550	0.1504	0.2762	0.4233	10.7072	0.7534
OISVM	0.9550	0.1444	0.4213	0.7639	13.8373	0.8400
		DOGMA ($B = 0.2$)				
PA-I	0.9813	0.0812	0.4680	0.7076	3.1330	0.8566
Perceptron	0.9641	0.1257	0.3023	0.4261	9.8340	0.7387
Projectron	0.9642	0.1254	0.3026	0.4264	9.7662	0.7391
Projectron++	0.9626	0.1289	0.2928	0.4105	10.5367	***0.7305***
RBP	0.9602	0.1346	0.2900	0.4169	11.2844	0.7287
Forgetron	0.9596	0.1353	0.2998	0.4413	12.7755	0.7285
OISVM	0.9718	0.1120	0.5014	0.8427	10.1921	**0.8619**
		DOGMA ($B = 0.3$)				
PA-I	0.9852	0.0686	0.5010	0.7636	2.2302	**0.8579**
Perceptron	0.9698	0.1114	0.3653	0.5447	6.0903	0.7729
Projectron	0.9706	0.1091	0.3492	0.5081	6.1235	0.7618
Projectron++	0.9737	0.1006	0.3575	0.5057	6.1688	0.7609
RBP	0.9641	0.1245	0.3662	0.5724	7.0734	0.7763
Forgetron	0.9666	0.1197	0.3648	0.5593	9.7172	***0.7592***
OISVM	0.9812	0.0840	0.5331	0.8627	9.3398	0.8521
		DOGMA ($B = 0.4$)				
PA-I	0.9911	0.0443	0.5415	0.8124	1.5328	0.8561
Perceptron	0.9790	0.0830	0.3972	0.5680	4.8458	0.7662
Projectron	0.9792	0.0817	0.3971	0.5657	4.8282	0.7656
Projectron++	0.9773	0.0884	0.3896	0.5633	7.0541	0.7537
RBP	0.9723	0.0989	0.3535	0.5025	9.0046	***0.7257***
Forgetron	0.9748	0.0913	0.3624	0.5063	6.1586	0.7411
OISVM	0.9889	0.0593	0.5632	0.8834	5.0767	**0.8596**
		DOGMA ($B = 0.5$)				
PA-I	0.9917	0.0444	0.5638	0.8647	1.2505	**0.8532**
Perceptron	0.9806	0.0849	0.4488	0.6805	3.6420	0.7859
Projectron	0.9793	0.0874	0.4442	0.6735	4.4825	0.7796
Projectron++	0.9830	0.0774	0.4274	0.6266	3.3437	0.7713
RBP	0.9702	0.1122	0.3998	0.6244	5.3916	***0.7604***
Forgetron	0.9688	0.1177	0.4062	0.6462	4.7573	0.7701
OISVM	0.9925	0.0429	0.5848	0.9059	6.2508	0.8405



(a) Entropy (E)

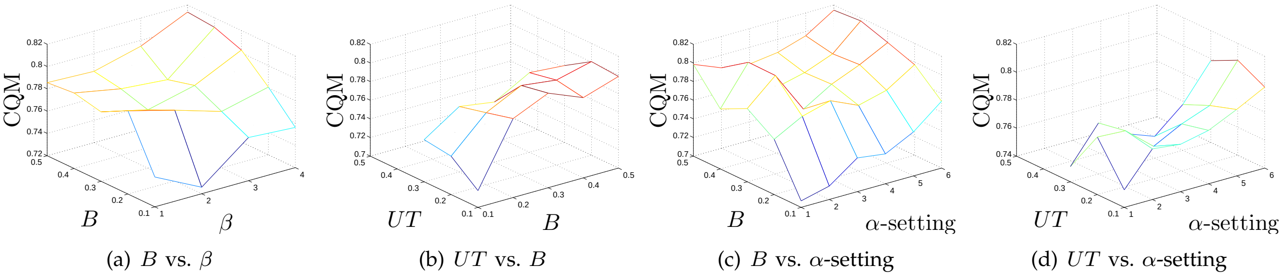
(b) Error Rate (ER)

(c) Purity (P)

(d) NMI

(e) Validity-measure (VM)

Fig. 11. Effect of UT and B on the accuracy using the synthetic social media dataset (SSMD)



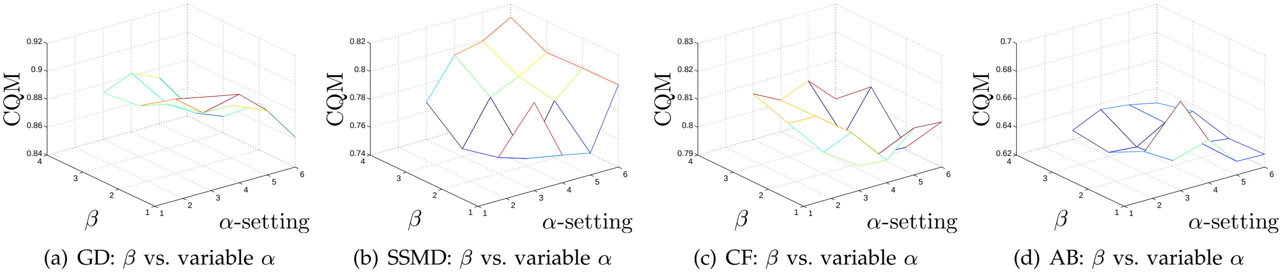
(a) B vs. β

(b) UT vs. B

(c) B vs. α-setting

(d) UT vs. α-setting

Fig. 12. CQM for various parameters using the synthetic social media dataset (SSMD)



(a) GD: β vs. variable α

(b) SSMD: β vs. variable α

(c) CF: β vs. variable α

(d) AB: β vs. variable α

Fig. 13. AOMPC (GD, SSMD, CF, AB): CQM for β and variable α

TABLE 8
DOGMA Algorithms (CF): Avg. results marking the **best** and ***worst*** CQM

	P	E	NMI	VM	ER	CQM
DOGMA ($B = 0.1$)						
PA-I	0.9276	0.1825	0.4159	0.7631	17.5100	**0.8214**
Perceptron	0.9273	0.1798	0.3930	0.7209	27.4159	0.7592
Projectron	0.9273	0.1798	0.3930	0.7209	27.4159	0.7592
Projectron++	0.9279	0.1780	0.3911	0.7137	28.4213	***0.7520***
RBP	0.9250	0.1868	0.3932	0.7277	26.0768	0.7679
Forgetron	0.9253	0.1843	0.3909	0.7206	27.3199	0.7596
OISVM	0.9154	0.2140	0.3844	0.7379	17.3795	0.8145
DOGMA ($B = 0.2$)						
PA-I	0.9346	0.1631	0.4316	0.7728	15.9354	**0.8122**
Perceptron	0.9384	0.1485	0.4117	0.7226	22.6292	0.7636
Projectron	0.9384	0.1485	0.4117	0.7226	22.6292	0.7636
Projectron++	0.9366	0.1537	0.4135	0.7304	21.5040	0.7716
RBP	0.9367	0.1517	0.4061	0.7141	23.7132	***0.7557***
Forgetron	0.9339	0.1610	0.4035	0.7198	23.6235	0.7578
OISVM	0.9250	0.1901	0.4079	0.7589	15.7008	0.8092
DOGMA ($B = 0.3$)						
PA-I	0.9443	0.1401	0.4595	0.8039	13.8672	**0.8118**
Perceptron	0.9416	0.1433	0.4161	0.7260	28.3164	0.7162
Projectron	0.9416	0.1433	0.4161	0.7260	28.3164	0.7162
Projectron++	0.9393	0.1476	0.4110	0.7198	26.6515	0.7227
RBP	0.9378	0.1533	0.4030	0.7128	33.5286	0.6862
Forgetron	0.9367	0.1560	0.4052	0.7180	29.8722	***0.7060***
OISVM	0.9335	0.1692	0.4321	0.7838	13.9892	0.8052
DOGMA ($B = 0.4$)						
PA-I	0.9499	0.1249	0.4790	0.8222	12.7396	**0.8030**
Perceptron	0.9449	0.1337	0.4376	0.7541	19.6706	0.7479
Projectron	0.9449	0.1337	0.4376	0.7541	19.6706	0.7479
Projectron++	0.9439	0.1367	0.4349	0.7518	19.4514	0.7483
RBP	0.9415	0.1431	0.4323	0.7546	21.3471	0.7396
Forgetron	0.9406	0.1445	0.4071	0.7117	28.5864	***0.6906***
OISVM	0.9424	0.1478	0.4592	0.8126	12.7191	0.8002
DOGMA ($B = 0.5$)						
PA-I	0.9544	0.1154	0.4927	0.8405	11.3371	**0.7955**
Perceptron	0.9511	0.1189	0.4500	0.7604	20.2214	0.7270
Projectron	0.9511	0.1189	0.4500	0.7604	20.2214	0.7270
Projectron++	0.9505	0.1211	0.4552	0.7728	18.2706	0.7405
RBP	0.9440	0.1377	0.4370	0.7559	20.7404	0.7231
Forgetron	0.9442	0.1347	0.4274	0.7353	24.1613	***0.6998***
OISVM	0.9482	0.1329	0.4777	0.8310	10.9356	0.7946

TABLE 9
DOGMA Algorithms (AB): Avg. results marking the **best** and ***worst*** CQM

	P	E	NMI	VM	ER	CQM
DOGMA ($B = 0.1$)						
PA-I	0.9070	0.2299	0.4041	0.6791	22.9801	**0.7688**
Perceptron	0.9058	0.2266	0.3899	0.6478	31.4818	0.7169
Projectron	0.9058	0.2266	0.3899	0.6478	31.4818	0.7169
Projectron++	0.9040	0.2302	0.3859	0.6440	32.6142	***0.7101***
RBP	0.9058	0.2266	0.3899	0.6478	31.4818	0.7169
Forgetron	0.9058	0.2266	0.3899	0.6478	31.4818	0.7169
OISVM	0.8955	0.2556	0.3833	0.6638	27.8570	0.7398
DOGMA ($B = 0.2$)						
PA-I	0.9231	0.1891	0.4420	0.7094	20.9924	**0.7678**
Perceptron	0.9112	0.2136	0.4050	0.6633	28.7431	0.7153
Projectron	0.9112	0.2136	0.4050	0.6633	28.7431	0.7153
Projectron++	0.9110	0.2135	0.4040	0.6617	29.3315	0.7119
RBP	0.9116	0.2134	0.4062	0.6652	28.3204	0.7180
Forgetron	0.9108	0.2150	0.4048	0.6643	29.6821	***0.7109***
OISVM	0.9133	0.2143	0.4242	0.7047	23.6358	0.7532
DOGMA ($B = 0.3$)						
PA-I	0.9317	0.1701	0.4692	0.7428	17.6217	**0.7747**
Perceptron	0.9150	0.2059	0.4146	0.6731	25.7420	0.7132
Projectron	0.9150	0.2059	0.4146	0.6731	25.7420	0.7132
Projectron++	0.9158	0.2050	0.4148	0.6739	25.6693	0.7138
RBP	0.9119	0.2134	0.4091	0.6707	27.3168	***0.7046***
Forgetron	0.9155	0.2050	0.4145	0.6739	25.9384	0.7125
OISVM	0.9248	0.1874	0.4552	0.7356	19.2589	0.7644
DOGMA ($B = 0.4$)						
PA-I	0.9419	0.1444	0.5016	0.7751	16.0927	**0.7721**
Perceptron	0.9248	0.1813	0.4431	0.7020	23.2030	0.7146
Projectron	0.9248	0.1813	0.4431	0.7020	23.2030	0.7146
Projectron++	0.9226	0.1870	0.4378	0.6990	24.3519	0.7079
RBP	0.9208	0.1890	0.4396	0.7010	24.6207	0.7072
Forgetron	0.9182	0.1979	0.4264	0.6870	24.4803	***0.7037***
OISVM	0.9386	0.1547	0.4945	0.7758	17.0144	0.7677
DOGMA ($B = 0.5$)						
PA-I	0.9500	0.1252	0.5272	0.8027	13.4840	0.7734
Perceptron	0.9293	0.1705	0.4584	0.7200	21.0540	0.7107
Projectron	0.9293	0.1705	0.4584	0.7200	21.0540	0.7107
Projectron++	0.9270	0.1776	0.4516	0.7147	21.7105	0.7059
RBP	0.9288	0.1714	0.4533	0.7111	22.2291	0.7022
Forgetron	0.9266	0.1772	0.4479	0.7086	22.5930	***0.6996***
OISVM	0.9492	0.1296	0.5259	0.8087	13.6702	**0.7743**